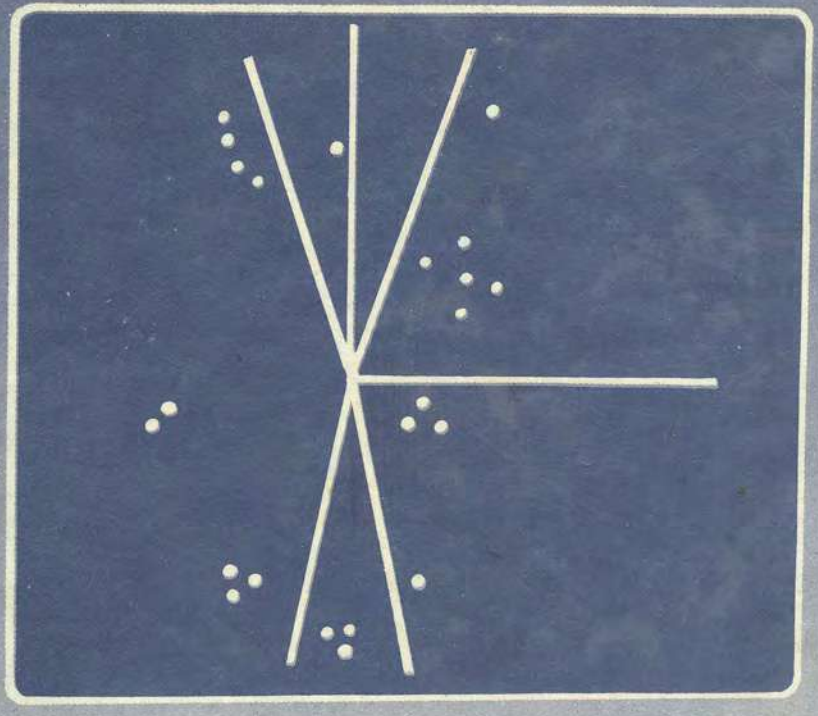
বহুচলক বিশ্লেষণ এবং এর প্রয়োগ



কেশব চন্দ্র ভূঞ্যাঁ

Web P

বহুচলক বিশ্লেষণ এবং এর প্রয়োগ

(Multivariate Analysis and its Applications)

ডঃ কেশব চন্দ্র ভুঞ্যাঁ
অধ্যাপক (অবসরপ্রাপ্ত)
পরিসংখ্যান বিভাগ
জাহাঙ্গীরনগর বিশ্ববিদ্যালয়

বাংলা একাডেমী ঢাকা

প্রথম প্রকাশ
জ্যৈষ্ঠ ১৪০৪/মে ১৯৯৭

বাং (১৯৯৬-৯৭ পাঠ্যপুস্তক : ভৌ ও প্রঃ ৬) ৩৫৪৫
মুদ্রণ সংখ্যা ১২৫০

পাণ্ডুলিপি প্রণয়ন ও মুদ্রণ তত্ত্বাবধান
ভৌতবিজ্ঞান ও প্রকৌশল উপবিভাগ
ভৌ ও প্র ১৬৬

প্রকাশক
আনিসুর রহমান
পরিচালক (ভারপ্রাপ্ত)
পাঠ্যপুস্তক বিভাগ
বাংলা একাডেমী, ঢাকা

মুদ্রক
মধুপুর প্রিন্টার্স
১৫/সি, আজিমপুর রোড
ঢাকা-১২০৫

প্রচ্ছদ
শওকাতুজ্জামান

মূল্য : ১৬০.০০

BOHUCHALAK BISHLESHAN EBONG ER PROYOG (Multivariate Analysis and its applications) By Dr. Keshab Chandra Bhuyan. Published by Anisur Rahman, Director (Incharge), Textbook Division, Bangla Academy, Dhaka, Bangladesh. First edition, May 1997. Price : Tk. 160.00

ISBN 984-07-3554-3

ভূমিকা

সামাজিক, অর্থনৈতিক, শিক্ষাসংক্রান্ত, মনোবিদ্যাগত, কৃষি-সংক্রান্ত এবং চিকিৎসাশাস্ত্রগত গবেষণায় একই বস্তু বা একক হতে বিভিন্নমুখী উপাত্ত সংগ্রহ করা একটি নিত্তনৈমিত্তিক ব্যাপার। সংগৃহীত উপাত্তের বিশ্লেষণের জন্য তাত্ত্বিক গবেষণারও অন্ত নেই। বিভিন্ন তাত্ত্বিক গবেষণার প্রয়োগও বর্তমানে সম্ভব হচ্ছে কম্পিউটার প্রোগ্রাম ব্যবহার করার কারণে। ফলে বহুচলক বিশ্লেষণ বর্তমান গবেষকদের নিকট বিশেষ করে ফলিত গবেষণাকারীদের নিকট একটি গুরুত্বপূর্ণ বিষয়।

বহুচলক বিশ্লেষণের ক্ষেত্রে যতো তাত্ত্বিক উন্নয়ন হয়েছে ঠিক ততোটা এর প্রয়োগ পরিসংখ্যানবিদদের দ্বারা হয় নি। মনোবিদ্যাগত গবেষণায় এর বহুল প্রয়োগ লক্ষ্য করা যায়। তবে সেক্ষেত্রে তাত্ত্বিক বিশ্লেষণ ও ফলিত বিশ্লেষণের মধ্যে একটি তফাৎ লক্ষ্য করা যায়। এর কারণ হলো তত্ত্বসমূহের প্রয়োগ উল্লেখপূর্বক বহুচলক বিশ্লেষণের বিস্তারিত ব্যাখ্যা খুব বেশি করা হয় নি। কিছু কিছু ক্ষেত্রে প্রয়োগ উল্লেখ থাকলেও বিশ্লেষিত ফলাফলের তাৎপর্য বা বিশ্লেষণে চলক-সমূহের ভূমিকা বা বিশ্লেষণ থেকে কোনো উপসংহারে পৌঁছার ব্যাপারে তেমন কিছু উল্লেখিত হয় নি। ফলে ছাত্রদের নিকট বা ফলিত গবেষকদের নিকট বিষয়টি সম্পর্কে তেমন আগ্রহ সৃষ্টি করা যায় নি।

এ গ্রন্থের সূচিপত্রের প্রতি লক্ষ্য করলেই বুঝা যাবে যে এখানে বহুচলক বিশ্লেষণের সবকিছু অন্তর্ভুক্ত না হলেও প্রয়োজনীয় তত্ত্ব-সমূহের ব্যাখ্যা দ্বিতীয় ও তৃতীয় অধ্যায়ে আলোচনা করা হয়েছে। চতুর্থ অধ্যায় থেকে নবম অধ্যায় পর্যন্ত প্রতিটি অধ্যায়ই গবেষণা কাজে প্রয়োগ করার মতো বিষয়বস্তু অন্তর্ভুক্ত করা হয়েছে। এ অধ্যায়-গুলোতে তাত্ত্বিক বিশ্লেষণ যেমন করা হয়েছে, তেমনি তত্ত্বসমূহের যথাযথ প্রয়োগ বাস্তব উপাত্তের ক্ষেত্রে দেখানো হয়েছে এবং বিশ্লেষিত ফলাফল থেকে একটি উপসংহারে পৌঁছাবার চেষ্টা করা হয়েছে। প্রধান উপাদান বিশ্লেষণ এবং উপাদান বিশ্লেষণের ক্ষেত্রে বাস্তব উপাত্তের বিশ্লেষণ করে ঐ দুটি বিশ্লেষণের মুখ্য উদ্দেশ্যকে ব্যাখ্যা করা হয়েছে। নির্ণায়ক বিশ্লেষণের ক্ষেত্রেও অনুরূপ ব্যাখ্যা দেয়া হয়েছে। এই তিনটি অধ্যায়ে বাস্তব উপাত্তের বিশ্লেষণ এবং বিশ্লেষিত ফলা-ফলের তাৎপর্য ব্যাখ্যা ফলিত গবেষণাকারীদের গবেষণার সহায়ক হবে আশা করি।

বিভিন্ন অধ্যায়ে যে সকল তত্ত্ব বা উপপাদ্য সংযোজিত হয়েছে সেগুলোর সবই প্রতিষ্ঠিত লেখকদের গ্রন্থ থেকে নেয়া হয়েছে। বহুচলক বিশ্লেষণের উপর যাঁরা গ্রন্থ লিখেছেন তাঁদের সকলের কাছেই আমি ঋণী। গ্রন্থটিতে তাত্ত্বিক দিক উপস্থাপনা করার ব্যাপারে মূলত Kshirsagar (1972) এবং Mardia et al (1988)-এর গ্রন্থের সাহায্য নেয়া হয়েছে বেশি করে। নিজস্ব গবেষণা কাজ এবং গ্যারিয়োনিস বিশ্ববিদ্যালয়ের তিনজন ছাত্রের এম. এসসি থিসিস-এর উপস্থাপনা করতে গিয়ে যে সব বিশ্লেষণ করতে হয়েছে তার ভিত্তিতেই তত্ত্বসমূহের বাস্তব প্রয়োগ দেখানো হয়েছে এবং ঐ সকল গবেষণালব্ধ ফলাফলের তাৎপর্য ব্যাখ্যা করা হয়েছে।

গ্রন্থটিতে যে সব উদাহরণ উল্লেখ করা হয়েছে সেগুলোর বিশ্লেষণের জন্য কম্পিউটার সহায়তা পাওয়া গিয়েছে পরিসংখ্যান বিভাগ, গ্যারিয়োনিস বিশ্ববিদ্যালয় থেকে। সেজন্য বিভাগীয় কর্মকর্তাদের নিকট আমি ঋণী। সহকর্মী মিঃ ইউসুফ গামাতি কম্পিউটার প্রোগ্রামের মাধ্যমে উপাত্ত বিশ্লেষণে আমাকে সবচেয়ে বেশি সহায়তা করেছেন। মিঃ গামাতিকে সে জন্য জানাই আন্তরিক ধন্যবাদ। অধ্যাপক আর. এন. শীল প্রাথমিক পাণ্ডুলিপিটি পড়েছেন এবং কিছু শব্দ পরিবর্তনের পক্ষে মতামত ব্যক্ত করেছেন। তাঁকেও আমি ধন্যবাদ জানাই।

পরিশেষে, গ্রন্থটি যাদের উদ্দেশ্যে লেখা হয়েছে সে সব ছাত্র-ছাত্রী এবং গবেষকগণ এর দ্বারা উপকৃত হলে আমার পরিশ্রম সার্থক হয়েছে বলে মনে করবো।

গ্রন্থকার

সূচিপত্র

প্রথম অধ্যায়

প্রাথমিক তথ্য
(Preliminary Information)

১.১ সূচনা (Introduction)

উদ্ভিদবিজ্ঞানী এবং জীববিজ্ঞানীদের গবেষণার একটি বিষয় হলো নতুন জীবজন্তু বা বৃক্ষসমূহের সন্ধান করা এবং নতুন আবিষ্কৃত জীবজন্তু বা চারাসমূহ কোন শ্রেণিভুক্ত সে সম্বন্ধে সিদ্ধান্ত গ্রহণ করা। একজন ডাক্তারের নিকট কোনো নতুন রোগী চিকিৎসার জন্য উপস্থিত হলে তিনি প্রথমে রোগীর রোগ নির্ণয়ের চেষ্টা করেন। ওষুধের কার্যকারিতা নির্ণয়ের জন্য গিনিপিগের উপর ওষুধ প্রয়োগ করা বহুল প্রচলিত। কৃষিক্ষেত্রে বিশেষ ফসলের জন্য জমির উপযুক্ততা নির্ণয় কৃষি বিজ্ঞানীদের গবেষণার বিষয়। শিক্ষাক্ষেত্রে ছাত্র ভর্তি বা চাকরির ক্ষেত্রে চাকরি প্রার্থীর উপযুক্ততা নির্ণয় বিশেষ বিশেষ কমিটির দায়িত্ব। আলোচিত প্রতিটি ক্ষেত্রেই সিদ্ধান্ত গ্রহণের জন্য গবেষক বা বিজ্ঞানী কেবল একটা চিহ্নিতকরণ উপাদানের উপর নির্ভর করতে পারেন না। যেমন, অন্য প্রাণী বধ করে মাংস খেলেই কোনো জন্তুকে বাঘ বলা যায় না। কোনো জন্তুর মধ্যে বাঘের সকল বৈশিষ্ট্য বিদ্যমান থাকলেই তাকে বাঘ বলে চিহ্নিত করা যাবে। গায়ের উত্তাপ পরিমাপ করে কোনো রোগীকে টাইফয়েডে আক্রান্ত বলে সিদ্ধান্ত গ্রহণ করা যায় না। টাইফয়েডের আরো লক্ষণ থাকতে হবে। গিনিপিগের উপর ওষুধ প্রয়োগ করে উহার কার্যকারিতা যাচাইয়ের গিনিপিগের শারীরিক পরিবর্তনের সকল চিহ্নই বিবেচনা করতে হবে। কৃষি জমির উপযুক্ততা যাচাই করার জন্য জমির উর্বরতা, জমিতে পটাশের পরিমাণ, মাটির প্রকারভেদ ইত্যাদি বিষয় বিবেচনা করতে হয়। ছাত্র ভর্তির ক্ষেত্রে অনেক সময় ছাত্রের পূর্ববর্তী পরীক্ষাসমূহের ফলাফল, ভর্তি পরীক্ষার ফলাফল বিবেচনা করে সিদ্ধান্ত গ্রহণ করা হয়। চাকরি ক্ষেত্রে প্রার্থী নির্বাচিত করার জন্য প্রার্থীর শিক্ষাগত যোগ্যতা, বয়স, শারীরিক অবস্থা ইত্যাদি বিষয় বিবেচনা করা হয়।

উপর্যুক্ত আলোচনা থেকে বুঝা যাচ্ছে যে, যে কোনো বস্তু সম্পর্কে কোনো সিদ্ধান্তে উপনীত হতে হলে ঐ বস্তুর একাধিক বৈশিষ্ট্য বা চলক পর্যালোচনা করতে হবে বা হয়। একাধিক চলকের মান একই বস্তু (object) থেকে পরিমাপ করা হয় বলে ঐ চলকসমূহ পরস্পর সম্পর্কিত (inter-related)। বৈজ্ঞানিক গবেষণার ক্ষেত্রে একটি বস্তু থেকে বিভিন্ন চলকের মান পরিমাপ করেও বস্তু

সম্পর্কে সিদ্ধান্ত নেয়া হয় না। সেক্ষেত্রে এক জাতীয় বস্তুর একটি নমুনা বা বহু নমুনা থেকে উপাত্ত সংগ্রহ করে সিদ্ধান্ত নেয়া যুক্তিগ্রাহ্য। সুতরাং বুঝা যাচ্ছে যে, এক জাতীয় বস্তুর এক বা একাধিক নমুনা থেকে সংগৃহীত বিভিন্ন চলকের উপাত্ত-ভিত্তিক বিশ্লেষণ বা ঐ চলকসমূহের মধ্যে কি ধরনের সম্পর্ক বিদ্যমান তা পর্যালোচনা করা কোনো কোনো গবেষকের গবেষণার বিষয়।

১.২ বহুচলক বিশ্লেষণ (Multivariate Analysis)

বহুচলক বিশ্লেষণ হলো এক বা একাধিক নমুনা বস্তুর প্রতিটি বস্তু হতে দুই বা ততোধিক চলকের উপাত্ত সংগ্রহ করে সংগৃহীত উপাত্তের পরিসংখ্যানিক বিশ্লেষণ। এই বিশ্লেষণের ক্ষেত্রে চলকসমূহের আন্তঃসম্পর্ক পর্যালোচনা করা হয়। এক্ষেত্রে এক চলক বা দ্বিচলক বিশ্লেষণের ন্যায় চলকের গড়, ভেদাঙ্ক বা চলকদ্বয়ের ক্ষেত্রে সংশ্লেষণ (correlation) বিশ্লেষণই যথেষ্ট নয়, বরং চলকসমূহের মধ্যে সম্পর্কের মাত্রা নির্ণয় করা হয়। উদাহরণ হিসেবে উল্লেখ করা যায় যে, পরিবার পরিকল্পনা গ্রহণ করার ক্ষেত্রে কোন কোন চলক বিশেষ বিশেষ দম্পতিকে বিশেষভাবে অনুপ্রাণিত করে তার পর্যালোচনা বা ভোক্তা কি কি কারণে একটি শিল্প পণ্য বিশেষভাবে ব্যবহার করবে তার পর্যালোচনা, বা কি কি বৈশিষ্ট্যের ভিত্তিতে একটি জন্তুকে বিশেষ শ্রেণিভুক্ত করা হবে তার পর্যালোচনা ইত্যাদি।

আগেই উল্লেখ করা হয়েছে যে, বহুচলক বিশ্লেষণে চলকসমূহের সম্পর্ক পর্যালোচনা করা হয়। কিন্তু চলকের সংখ্যা বেশি হলে সেগুলোর মধ্যে সংশ্লেষণের সংখ্যাও বেশি হয়। ফলে বিশ্লেষণ জটিল হয়। এরূপ ক্ষেত্রে মূল তথ্য যতটা সম্ভব বজায় রেখে অল্প চলক ব্যবহার করে বিশ্লেষণ করার পদ্ধতি পর্যালোচনাও বহুচলক বিশ্লেষণের একটি অঙ্গ। এখানে অল্প চলক ব্যবহার করার পদ্ধতিকে উপাত্ত সংক্ষেপণ (data reduction) পদ্ধতি বলা হয়। উপাত্ত সংক্ষেপিত করার দুটি পদ্ধতি আলোচনা করা হবে। এ দুটির প্রথমটি হলো প্রধান উপাদান বিশ্লেষণ (principal component analysis) এবং দ্বিতীয়টি হলো উপাদান বিশ্লেষণ (factor analysis)। যথাক্রমে চতুর্থ এবং পঞ্চম অধ্যায়ে এ দুটি বিশ্লেষণ পদ্ধতি আলোচনা করা হবে। কোনো নমুনা থেকে প্রাপ্ত বস্তুসমূহ (objects, individuals) পর্যালোচনা করে যে চলকসমূহ বিশ্লেষণের জন্য নির্বাচিত করা হয় সেগুলোকে দুটি গুচ্ছে (set) বিভক্ত করা যায়। এই বিশ্লেষণ পদ্ধতিকে কানুনী সংশ্লেষণ বিশ্লেষণ (canonical correlation analysis) বলা হয়। ষষ্ঠ অধ্যায়ে এই বিশ্লেষণ পদ্ধতি আলোচনা করা হবে। নমুনা বস্তুসমূহ থেকে পরিমাপকৃত অনপেক্ষ চলকের ভিত্তিতে বস্তুকে বা বস্তুসমূহকে বিসদৃশ গ্রুপে ভাগ করা বা একটি নতুন বস্তু কোন গ্রুপভুক্ত হবে তা নির্ণয় করাও বহুচলক বিশ্লেষণের একটি অঙ্গ। এই বিশ্লেষণ পদ্ধতিকে বলা হয় নির্ণায়ক বিশ্লেষণ

(discriminant analysis)। এটি সপ্তম অধ্যায়ে আলোচনা করা হবে। যেহেতু একই বস্তুর অনেকগুলো চলক পরিমাপ করা হয়, সেই বহুচলকের ভিত্তিতে বস্তুসমূহকে k গ্রুপে বিভক্ত করা যেতে পারে। বস্তুসমূহকে দুই-এর অধিক গ্রুপে বিভক্ত করার পদ্ধতি বহুচলক ভেদাঙ্ক বিশ্লেষণ (multivariate analysis of variance) নামে পরিচিত। এটি অষ্টম অধ্যায়ে আলোচনা করা হবে। এ হলো বহুচলক বিশ্লেষণের একটি সংক্ষিপ্ত ধারণা। এ সম্পর্কিত আরো কিছু বিশ্লেষণ পদ্ধতি ভিন্ন অধ্যায়ে আলোচনা করা হবে।

১.৩ উপাত্ত (Data)

আগেই আলোচনা করা হয়েছে যে, বহুচলক বিশ্লেষণের ক্ষেত্রে একই বস্তু হতে বিভিন্ন চলকের মান পরিমাপ করা হয়। ধরা যাক, কোনো গণসমষ্টি হতে n আকারের একটি নমুনা চয়ন করা হয়েছে এবং নমুনার প্রতিটি বস্তু হতে p চলকের মান পরিমাপ করা হয়েছে। আরো ধরা যাক, প্রতিটি চলকের মান k বার বা k ঘটনার পরিপ্রেক্ষিতে (occasion) পরিমাপ করা হয়েছে। ধরা যাক x_{ijl} [$i = 1, 2, \cdots, n$; $j = 1, 2, \cdots, p$; $l = 1, 2, \cdots, k$] হলো i-তম বস্তুর j-তম চলকের l-তম বার পরিমাপ করা মান। প্রাপ্ত উপাত্ত x_{ijl} এর সকল মানকে নিম্নরূপভাবে সাজিয়ে লেখা যায় :

বস্তু	পুনরায়ন	চলক					
		1	2	⋯	j	⋯	p
1	1	x_{111}	x_{121}	⋯	x_{1j1}	⋯	x_{1p1}
	2	x_{112}	x_{122}	⋯	x_{1j2}	⋯	x_{1p2}
	⋮	⋯	⋯		⋯		⋯
	i	x_{11i}	x_{12i}	⋯	x_{1ji}	⋯	x_{1pi}
	⋮	⋯	⋯		⋯	⋯	⋯
	k	x_{11k}	x_{12k}	⋯	x_{1jk}	⋯	x_{1pk}
2	1	x_{211}	x_{221}	⋯	x_{2j1}	⋯	x_{2p1}
	2	x_{212}	x_{222}	⋯	x_{2j2}	⋯	x_{2p2}
	⋮	⋯	⋯	⋯	⋯	⋯	⋯
	i	x_{21i}	x_{22i}	⋯	x_{2ji}	⋯	x_{2pi}
	⋮	⋮	⋮		⋮	⋮	⋮
	k	x_{21k}	x_{22k}	⋯	x_{2jk}	⋯	x_{2pk}

...	...	...	...	...	...	...	...	...	...	...	...	...	...
		1		x_{i11}	x_{i21}	...	...	x_{ij1}	...	...	x_{ip1}
		2		x_{i12}	x_{i22}	...	...	x_{ij2}	...	...	x_{ip2}
i		⋮	...	...	...	...	...	...	...	...	...
		i		x_{i1i}	x_{i2i}	...	...	x_{iji}	...	...	x_{ipi}
		⋮		⋮	⋮	...	...	⋮	⋮	⋮	⋮
		k		x_{i1k}	x_{i2k}	...	...	x_{ijk}	...	...	x_{ipk}
...	...	...	...	...	...	...	...	...	...	...	...
		1		x_{n11}	x_{n21}	...	...	x_{nj1}	...	...	x_{np1}
		2		x_{n12}	x_{n22}	...	...	x_{nj2}	...	...	x_{np2}
		⋮		...	...	...	...	...	...	...	...
n		1		x_{n1i}	x_{n2i}	...	...	x_{nji}	...	...	x_{npi}
		⋮		⋮	⋮	...	...	⋮	⋮	⋮	⋮
		k		x_{n1k}	x_{n2k}	...	...	x_{njk}	...	...	x_{npk}

উপরিউক্তভাবে x_{iji} এর মানসমূহ সাজিয়ে লেখা হলে যে ম্যাট্রিক্স-এর উদ্ভব হয় তাকে সুপার ম্যাট্রিক্স [Cattell (1952)] বলা হয়। বাস্তবক্ষেত্রে $k=1$ হলেও বিশ্লেষণ করা যায়। সেক্ষেত্রে প্রাপ্ত উপাত্তসমূহকে সারি এবং স্তম্ভে নিম্নরূপভাবে সাজানো যেতে পারে :

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & \cdots & x_{1j} & \cdots & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & \cdots & x_{2j} & \cdots & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ x_{i1} & x_{i2} & \cdots & \cdots & x_{ij} & \cdots & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & \cdots & x_{nj} & \cdots & \cdots & x_{np} \end{bmatrix}$$

এই X-কে বলা হয় $(n \times p)$ উপাত্ত ম্যাট্রিক্স। একে অন্যভাবেও লেখা যায়।

$$X=(x_{ij})=\begin{bmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{bmatrix}=[X_{(1)}\, X_{(2)}\, \cdots\, X_{(p)}]$$

এখানে $X_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}$, $X_{(j)} = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix}$

X_i হলো i-তম বস্তুর p চলকের মানভিত্তিক ভেক্টর এবং $X_{(j)}$ হলো n বস্তুর j-তম চলকের মানভিত্তিক ভেক্টর।

উদাহরণ হিসেবে Ghot Sultan Diary Project, Libya [Gamati (1992)] থেকে প্রাপ্ত দিনে দুইবার এবং তিনবার দোহন করা ২৪টি গরুর দৈনিক দুগ্ধ উৎপাদনের পরিমাণ ও সংশ্লিষ্ট কিছু চলকের মানের কথা উল্লেখ করা যেতে পারে।

উদাহরণ ১.১ ঃ দিনে দুইবার দোহন করা গরুর (C_2) দুধ উৎপাদনের পরিমাণ ও সংশ্লিষ্ট চলকের তথ্য।

ক্রমিক সংখ্যা	দৈনিক দুধ উৎপাদনের পরিমাণ A (kg)	গরুর ওজন, প্রাথমিক B (kg)	দুধ উৎপাদনকাল শেষ হওয়ার পর C(kg)	উৎপাদনকাল শেষ হওয়ার পর শারীরিক অবস্থা (BCS) D	মেসটিটিস E	বাছুর প্রসবের সময় বয়স F (দিন)
1	30.6	810	745	4.0	0	4698
2	17.1	860	920	5.0	1	1714
3	31.8	780	730	4.5	0	1623
4	23.9	705	710	4.5	0	1527
5	26.8	770	720	4.5	0	1685
6	34.6	695	700	3.5	0	1750
7	32.1	780	800	4.0	0	1678
8	23.5	740	740	3.5	0	1718
9	30.7	825	825	4.5	0	1699
10	21.3	790	790	3.5	0	1770
11	25.8	780	730	4.5	0	1682

12	31.4	920	830	3.0	1	1402
13	29.8	850	780	4.5	0	1724
14	35.7	755	740	3.5	0	1654
15	29.7	845	805	3.5	0	1651
16	27.6	740	780	4.5	0	1703
17	28.0	840	875	4.5	0	1745
18	26.9	765	785	5.0	0	1461
19	20.6	770	845	5.0	0	1558
20	28.9	805	805	4.0	0	1724
21	23.3	740	860	4.5	0	1606
22	28.4	670	735	4.5	0	1473
23	25.1	865	815	4.0	1	1609
24	31.8	715	705	3.5	0	1497
25	19.9	780	900	5.0	0	1500
26	23.0	720	810	4.5	0	1524
27	32.8	675	645	4.0	0	1610
28	26.5	795	790	4.0	0	1762

দিনে তিনবার দোহন করা গরুর (C_2) দুধ উৎপাদনের পরিমাণ ও সংশ্লিষ্ট চলকের তথ্য।

ক্রমিক সংখ্যা	A	B	C	D	E	F
1	34.6	810	730	3.5	0	1706
2	27.7	830	755	4.0	0	1655
3	29.2	740	710	3.5	0	1589

4	25.3	750	780	4.5	0	1541
5	27.6	800	845	5.0	1	1682
6	37.9	725	690	4.0	0	1437
7	32.6	760	800	4.5	1	1542
8	32.0	790	805	3.5	1	1669
9	30.7	910	835	5.0	0	1738
10	29.6	730	785	4.5	0	1760
11	38.3	820	805	4.0	0	1406
12	32.9	850	850	4.5	0	1660
13	30.8	775	770	4.0	0	1642
14	32.2	845	835	4.5	0	1782
15	32.9	890	850	4.5	1	1675
16	28.1	780	755	4.5	1	1748
17	33.9	755	715	3.5	0	1727
18	28.6	800	835	4.5	1	1769
19	28.1	775	800	4.5	0	1481
20	35.9	860	815	4.0	1	1784
21	34.8	670	630	3.5	1	1665
22	40.3	710	730	4.0	0	1679
23	30.9	770	785	4.5	0	1758
24	34.4	690	680	3.5	0	1711
25	19.8	775	840	5.0	1	1556
26	25.8	680	755	4.0	1	1524
27	37.3	695	630	2.5	0	1577
28	32.4	750	740	4.0	1	1524

উপর্যুক্ত উদাহরণের উপাত্তকে $k=2$ ঘটনার পরিপ্রেক্ষিতে (occasions) পরিমাপ করাকে উপাত্ত বলা যায়। সেক্ষেত্রে $n=28$ এবং $p=6$। মোট তথ্যের পরিমাণ হলো $npk=336$। আবার শুধু C_2 বা C_3 এর উপাত্ত বিবেচনা করা হলে মোট উপাত্তের পরিমাণ হয় $np=168$। দেখা যাচ্ছে যে, তথ্যের পরিমাণ খুব বেশি। এটি আরো বেশি হবে যদি n বা p বা উভয়ে বড় হয়। সে ক্ষেত্রে বিশ্লেষণের সুবিধার জন্য বা উপাত্ত সম্পর্কে একটি সহজ ধারণা লাভ করার জন্য n বস্তুর প্রতিটি চলকের গড়, ভেদাঙ্ক এবং সহ-ভেদাঙ্ক নির্ণয় করা যেতে পারে।

১.৪ উপাত্ত সারাংশ (Data Summary)

উপাত্ত সম্পর্কে প্রাথমিকভাবে সহজ একটি ধারণা লাভ করার জন্য সেগুলোর গড়, ভেদাঙ্ক, সহ-ভেদাঙ্ক, সংশ্লেষাঙ্ক নির্ণয় করা যেতে পারে।

গড় ঃ j-তম চলকের গড় হলো

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \; ; \qquad j = 1, 2, \cdots, p$$

সব চলকের গড় নির্ণয় করে সেগুলোকে একটি ভেক্টরের মাধ্যমে প্রকাশ করা যায়। যেমন,

$$\bar{X} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

এই $\bar{X}$-কে নমুনা গড় ভেক্টর বলা হয়। ১.১ উদাহরণের ক্ষেত্রে c_2 এবং c_3 এর উপাত্তের জন্য এরূপ দুটি নমুনা ভেক্টর হলো যথাক্রমে ঃ

$$\bar{X}_2 = \begin{bmatrix} 27.41 \\ 778.04 \\ 782.68 \\ 4.20 \\ 0.11 \\ 1633.82 \end{bmatrix} \qquad \bar{X}_3 = \begin{bmatrix} 31.59 \\ 776.25 \\ 769.82 \\ 4.13 \\ 0.39 \\ 1642.39 \end{bmatrix}$$

নমুনা ভেক্টরকে ম্যাট্রিক্স চিহ্নের মাধ্যমেও প্রকাশ করা যায়। যেমন :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n} X'1$$

এখানে $1 = [1\ 1\ \cdots\ 1]'_{1\times n}$। এই $\bar{X}$-কে বলা হয় centroid।

j-তম চলকের নমুনা ভেদাঙ্ক এবং j-তম ও k-তম চলকের সহ-ভেদাঙ্ককে লেখা যায়, যথাক্রমে

$$S_{jj} = \frac{1}{n}\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 = S_j^2, \quad j = 1, 2, \cdots, p$$

$$S_{jk} = \frac{1}{n}\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - \bar{x}_j\bar{x}_k$$

এই S_{jk} এর সকল মান (j ও k এর সকল মানের জন্য) ম্যাট্রিক্স আকারে সাজিয়ে লেখা যায়

$$S = (S_{jk}) = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})' = \frac{1}{n}M - \bar{X}\bar{X}'$$

$$= \frac{1}{n}X'X - \bar{X}\bar{X}' = \frac{1}{n}\left(X'X - \frac{1}{n}X'11'X\right)$$

$$= \frac{1}{n}X'HX, \quad \text{এখানে } M = X'X$$

এখানে $H = I - \frac{1}{n}11'$; এখানে I হলো $n \times n$ আকারের ইউনিট ম্যাট্রিক্স এবং S হলো নমুনা। এই নমুনা সহ-ভেদাঙ্ক ম্যাট্রিক্স বহুচলক বিস্তার পরিমাপ করার জন্য ব্যবহৃত হয়। বিস্তার পরিমাপ করার জন্য দুটি তথ্যজমান হলো, (১) $|S|$, জেনারালাইজড ভেদাঙ্ক এবং (২) মোট বিস্তার trs।

১.১ উদাহরণের ক্ষেত্রে C_2 ও C_3 এর উপাত্তের ভিত্তিতে এরূপ দুটি নমুনা সহ-ভেদাঙ্ক ম্যাট্রিক্স হলো, যথাক্রমে :

$$S_2 = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{bmatrix} 20.76 & -27.15 & -165.38 & -1.30 & -0.31 & 23.69 \\ & 3548.82 & 2286.51 & -3.01 & 11.10 & 981.08 \\ & & 4006.22 & 12.60 & 7.75 & -48.63 \\ & & & 0.29 & -0.02 & -6.41 \\ & & & & 0.10 & -6.30 \\ & & & & & 10312.65 \end{bmatrix} \end{array}$$

ও

$$S_3 = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{bmatrix} 19.21 & -20.24 & -112.41 & -1.33 & -0.61 & -1.06 \\ & 3660.04 & 2847.54 & 16.90 & 0.58 & 1950.76 \\ & & 3854.43 & 27.79 & 7.21 & 1159.36 \\ & & & 0.31 & 0.06 & 6.67 \\ & & & & 0.24 & 2.56 \\ & & & & & 11133.31 \end{bmatrix} \end{array}$$

লক্ষ্য করা গিয়েছে যে, $S = n^{-1} X' HX$ । কিন্তু H প্রতিসম (symmetric) আইডেমপোটেন্ট ম্যাট্রিক্স হওয়ার কারণে যে কোনো p-ভেক্টর a এর জন্য পাওয়া যায়

$$a' Sa = \frac{1}{n} a' X' H' HXa = \frac{1}{n} y' y \geq 0$$

এখানে $y = HXa$ । সুতরাং ভেদাঙ্ক ম্যাট্রিক্স S হলো ধনাত্মক সেমি-ডেফিনিট। এটি ধনাত্মক ডেফিনিট ম্যাট্রিক্স হতে পারে যদি $n \geq p+1$ হয় এবং চলকসমূহ অবিচ্ছিন্ন হয় ।

আলোচিত সহ-ভেদাঙ্ক ম্যাট্রিক্স-এর ভিত্তিতে চলকসমূহের সরল সংশ্লেষাঙ্ক ম্যাট্রিক্স R নির্ণয় করা যায় । যেখানে

$$R = (r_{jk})$$

এবং $r_{jk} = S_{jk}/S_jS_k$ ($j \neq k$, $|r_{jk}| \leqslant 1$ এবং $r_{jj} = 1$)। চলকসমূহ সংশ্লিষ্ট হলে $R \geqslant 0$, নতুবা $R = 1$। এই R-কে লেখা যায়

$$R = D^{-1}SD^{-1}$$

বা $$S = DRD, \quad \text{যেখানে } D = \text{diag}(S_j)$$

১.১ উদাহরণের ক্ষেত্রে C_2 ও C_3 এর উপাত্তের জন্য এরূপ দুটি নমুনা সংশ্লেষাঙ্ক ম্যাট্রিক্স (sample correlation matrix) হলো, যথাক্রমে

$$R_2 = \begin{bmatrix} 1.000 & -0.100 & -0.573 & -0.529 & -0.219 & 0.051 \\ & 1.000 & 0.606 & -0.093 & 0.603 & 0.162 \\ & & 1.000 & 0.369 & 0.396 & -0.008 \\ & & & 1.000 & -0.126 & -0.117 \\ & & & & 1.000 & -0.201 \\ & & & & & 1.000 \end{bmatrix}$$

এবং

$$R_3 = \begin{bmatrix} 1.000 & -0.076 & -0.413 & -0.541 & -0.284 & -0.002 \\ & 1.000 & 0.758 & 0.498 & 0.020 & 0.306 \\ & & 1.000 & 0.803 & 0.238 & 0.177 \\ & & & 1.000 & 0.212 & 0.113 \\ & & & & 1.000 & 0.050 \\ & & & & & 1.000 \end{bmatrix}$$

উপরের আলোচনায় S ম্যাট্রিক্সকে লেখা হয়েছে $S = n^{-1}X'HX$। কিন্তু নমুনা ভেদাঙ্ক ও সহ-ভেদাঙ্ক নির্ণয় করার সময় বর্গসমষ্টি বা গুণন-সমষ্টিকে n এর পরিবর্তে $(n-1)$ দ্বারা ভাগ করা হয়। অর্থাৎ নমুনা সহ-ভেদাঙ্ক ম্যাট্রিক্সকে লেখা যেতে পারে

$$S_s = (n-1)^{-1}X'HX = n(n-1)^{-1}S$$

১.৫ উপাত্ত পরিমাপের ধরন (Types of Measurement of Data)

১.৩ অনুচ্ছেদে উদাহরণসহ উপাত্ত সম্পর্কে আলোচনা করা হয়েছে। লক্ষ্য করা গিয়েছে যে, ১.১ উদাহরণে সকল চলকের পরিমাপ একইরূপ নয়। যেমন, চলক

E এর মান ধরা হয়েছে 0 বা 1, যথাক্রমে মেসটিটিস-এর অনুপস্থিতি বা উপস্থিতি অনুসারে। আবার, D চলকের মান পরিমাপ করা হয়েছে কতকগুলো গুণবাচক চলকের ভিত্তিতে। অপরদিকে চলক A, B, C এবং F এর মান পরিমাপ করা হয়েছে স্বাভাবিক পদ্ধতিতে। দেখা যাচ্ছে যে চলকের মান পরিমাপ পদ্ধতি একই বস্তুর সকল চলকের জন্য এক নয়। কিন্তু বহুচলক বিশ্লেষণের বিভিন্ন পদ্ধতির প্রয়োগ চলকের বিন্যাসের সাথে সম্পর্কিত—এটি পরবর্তী অনুচ্ছেদসমূহে লক্ষ্য করা যাবে। সাধারণত চলকের বিন্যাস বহুচলক পরিমিত বিন্যাস (multivariate normal distribution) অনুমান করা হয়। কারণ বিশ্লেষণের সাথে জড়িত যাচাই পদ্ধতি (test procedure) বহুচলক পরিমিত বিন্যাসের উপর নির্ভরশীল। কিন্তু চলকের পরিমাপ যদি নমিনাল (nominal) বা অর্ডিনাল (ordinal) পদ্ধতিতে পরিমাপ করা হয়, তাহলে ঐ চলক পরিমিত বিন্যাস অনুসরণ করবে তা হলফ করে বলা যায় না। ১.১ উদাহরণের ক্ষেত্রে চলক E এর মান পরিমাপ করা হয়েছে নমিনাল পদ্ধতিতে। চলক A, B, C এবং F এর মান পরিমাপ করা হয়েছে ব্যাপ্তি-মাপনী (interval scale) পদ্ধতিতে এবং D এর মান পরিমাপ করা হয়েছে স্কোরিং (scoring) পদ্ধতিতে। এখানে চলক A, B, C, D এবং F এর বিন্যাস পরিমিত বিন্যাস হতে পারে। অবশ্য চলকগুলো ভিন্নভাবে পরিমিত বিন্যাস অনুসরণ করলেই যে তাদের যুগ্ম বিন্যাস বহুচলক পরিমিত বিন্যাস হবে এমন কথা বলা যায় না। সে যাই হোক, বহুচলক বিশ্লেষণের ক্ষেত্রে বিচ্ছিন্ন এবং অবিচ্ছিন্ন উভয় ধরনের চলকের ব্যবহার লক্ষ্য করা যায় [Mardia et al (1988)]। কাজেই চলকের মান পরিমাপ করার জন্য ব্যাপ্তি-মাপনী পদ্ধতি, নমিনাল এবং অর্ডিনাল পদ্ধতির ব্যবহার লক্ষ্য করা যায় [Dillon and Goldstein (1984)]।

১.৬ উপাত্ত পরিবর্তন (Data Transformation)

বহুচলক বিশ্লেষণে একই বস্তুর অনেকগুলো বৈশিষ্ট্য পর্যালোচনা করা হয় বলে চলকের সংখ্যা বৃদ্ধি পায় এবং বর্ধিত চলক ব্যবহার করে অনেক সময় সিদ্ধান্তে উপনীত হওয়া কষ্টকর হয়। সেক্ষেত্রে চলকের একটি রৈখিক সমাবেশ (linear combination) ব্যবহার করে অধিক অর্থবহ সিদ্ধান্ত গ্রহণ পদ্ধতি পর্যালোচনা করা বহুচলক বিশ্লেষণের একটি অঙ্গ। রৈখিক সমাবেশ ব্যবহার করার কারণে চলকের সংখ্যা কম হতে পারে এবং বিশ্লেষণ সহজতর হতে পারে। চলকের রৈখিক সমাবেশ উপাত্ত পরিবর্তনের একটি অংশ। তাছাড়া অনেক সময় বিভিন্ন চলকের মান বিভিন্ন এককে (unit) পরিমাপ করা হয় বলে তাদের ভেদাঙ্কের পরিমাণে বিভিন্নতা লক্ষ্য করা যায় কিন্তু উপাত্ত পরিবর্তন প্রক্রিয়ার মাধ্যমে সব চলককে এক ভেদাঙ্কবিশিষ্ট চলকে প্রকাশ করা যায়। ফলে উপাত্তের ভেদ সম্পর্কে একটি সহজ ধারণা লাভ করা যায়। বর্তমান অনুচ্ছেদে কিছু উপাত্ত পরিবর্তন পদ্ধতি আলোচনা করা হবে। প্রথমে চলকগুলোর একটি রৈখিক সমাবেশ বিবেচনা করা যাক।

মনে করি $x_1, x_2, \cdots, x_p$ চলকের একটি রৈখিক সমাবেশ হলো

$$Y_i = a_1x_{i1} + a_2x_{i2} + \cdots + a_px_{ip} \; ; \quad i = 1, 2, \cdots, n ;$$

এখানে $a_1, a_2, \cdots, a_p$ হলো জানা মান। নতুন চলক Y_i এর গড় হলো

$$\bar{Y} = \frac{1}{n} a' \sum_{i=1}^{n} X_i = a' \bar{X}$$

এখানে $a' = [a_1\ a_2 \cdots a_p]$। চলক Y_i এর ভেদাঙ্ক হলো

$$S_y^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \frac{1}{n} \sum_{i=1}^{n} a'(X_i - \bar{X})(X_i - \bar{X})'a = a'Sa$$

উপরে আলোচিত চলকের পরিবর্তন হলো এক মাত্রার পরিবর্তন। বাস্তবে বহু মাত্রার পরিবর্তন করা যেতে পারে। ধরা যাক, বহু মাত্রার পরিবর্তনের জন্য পরিবর্তন সহগ ভেক্টর a এর পরিবর্তে A হলো জানা মানসমূহের একটি $(q \times p)$ ম্যাট্রিক্স এবং b হলো একটি q-ভেক্টর। তাহলে পরিবর্তিত চলক-এর ভিত্তিতে উপাত্ত ম্যাট্রিক্সকে লেখা যায়

$$Y = XA' + 1b'$$

এখান থেকে লেখা যায়

$$\bar{Y} = A\bar{X} + b$$

এবং $$S_y^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})(Y_i - \bar{Y})' = ASA'$$

যদি A ম্যাট্রিক্স নন্-সিঙ্গুলার হয় এবং $q = p$ হয়, তাহলে

$$S = A^{-1} S_y (A')^{-1}$$

এতক্ষণ একটি সাধারণ পরিবর্তন আলোচনা করা হয়েছে। চলকসমূহকে বিশেষ বিশেষ পরিবর্তনের মাধ্যমেও প্রকাশ করা যায়। এরূপ কতকগুলো পরিবর্তন হলো : (১) স্কেলিং পরিবর্তন, (২) Mahalanobis পরিবর্তন এবং (৩) মূল উপাদান পরিবর্তন। এ সকল পরিবর্তনের জন্য চলকের আদি বিন্দু পরিবর্তনও বিবেচনা করা হবে।

স্কেলিং পরিবর্তন (Scaling Transformation) : এই পরিবর্তনের উদ্দেশ্য হলো চলকসমূহকে এক ভেদাঙ্কবিশিষ্ট চলকে পরিবর্তন করা। ধরা যাক $D = \text{diag}(S_j)$। তাহলে D প্রয়োগ করে পরিবর্তিত চলককে লেখা যায়

$$Y_i = D^{-1}(X_i - \bar{X}) \; ; \quad i = 1, 2, \cdots, n$$

Mahalanobis পরিবর্তন (Mahalanobis Transformation) : এই পরিবর্তনের মাধ্যমে নতুন চলক হলো

$$Z_i = S^{-1/2}(X_i - \bar{X}) \; ; \quad i = 1, 2, \cdots, n$$

এখানে শর্ত হলো $S > 0$ হবে। তাহলে S^{-1} এর একক প্রতিসম ধনাত্মক ডেফিনিট বর্গমূল (positive definite square root) $S^{-1/2}$ পাওয়া যাবে। এই পরিবর্তনের ফলে $S_z = I$ হবে এবং চলকসমূহের মধ্যে কোনো সংশ্লেষণ থাকবে না। এটি প্রতিটি চলকের ভেদাঙ্ককে আদর্শায়িত করে।

প্রধান উপাদান পরিবর্তন (Principal Component Transformation) : Spectral decomposition উপপাদ্যের মাধ্যমে [যে কোনো প্রতিসম ম্যাট্রিক্স $S(p \times p)$-কে লেখা যায় $S = \Gamma \Lambda \Gamma'$, যেখানে $\Lambda = \text{diag}(\lambda_i)$ এবং Γ হলো সমকৌণিক ম্যাট্রিক্স। এখানে λ_i হলো S ম্যাট্রিক্স-এর আইগেন মান] S-কে লেখা যায় $S = \Gamma \Lambda \Gamma'$, যেখানে $\Lambda = \text{diag}(\lambda_j)$, Γ হলো একটি সমকৌণিক ম্যাট্রিক্স এবং λ_j হলো S ম্যাট্রিক্স এর আইগেন মান। ধরা যাক $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geqslant 0$। তাহলে প্রধান উপাদান পরিবর্তনের সংজ্ঞা হলো

$$W_i = \Gamma'(X_i - \bar{X}), \quad i = 1, 2, \cdots, n$$

এই পরিবর্তনের ফলে $S_w = \Gamma' S \Gamma = \Lambda$। এখানে W ম্যাট্রিক্স এর স্তম্ভসমূহ হলো প্রধান উপাদানসমূহ। প্রধান উপাদানসমূহ অসংশ্লেষিত এবং তাদের ভেদাঙ্ক হলো $\lambda_1, \lambda_2, \cdots, \lambda_p$। সুতরাং, বহুচলক বিস্তার পরিমাপ করার তথ্যজমান হলো

$$|S| = |\Lambda| = \prod_{j=1}^{p} \lambda_j \text{ এবং } \text{tr}S = \text{tr}\Lambda = \Sigma\lambda_j$$

উপরে আলোচিত স্কেলিং পরিবর্তন ও Mahalanobis পরিবর্তন Bhuyan and Nair (1995) এর কাজের আংশিক উপাত্তের ক্ষেত্রে প্রয়োগ করা যাক।

উদাহরণ ১.২ : ১৯৯৩ সনের ফেব্রুয়ারি—মার্চ মাসে লিবিয়ার পূর্বাঞ্চল হতে সংগৃহীত চার প্রকার Slugs এর বিভিন্ন অঙ্গ-প্রত্যঙ্গের পরিমাপ করা হয় [Bhuyan and Nair (1995)]। নিচে প্রতি প্রকারের পনেরটি Slugs এর দৈহিক ওজন (Body weight, W) এবং দৈহিক দৈর্ঘ্য (Body lenghth, L) দেয়া হলো।

Slugs :	M. Rusticus		M. Gagates		M. Tenellus		M. Sowerbye	
	L(mm)	W(gms)	L(mm)	W(gms)	L(mm)	W(gms)	L(mm)	W(gms)
	35	1.3	35	2.5	41	1.4	32	0.6
	35	4.0	40	1.7	42	1.4	40	2.4
	38	3.2	40	2.0	44	1.8	41	1.5
	35	1.0	40	2.0	44	1.4	41	2.3
	39	1.4	40	2.5	45	1.2	45	1.9
	40	1.8	40	3.0	45	1.8	45	1.9
	40	2.3	40	3.7	45	1.7	52	2.8
	40	2.1	40	4.2	46	1.9	55	2.4
	40	2.5	40	4.8	49	2.0	55	3.4
	40	3.5	40	3.4	49	2.3	55	3.9
	40	2.1	41	3.3	50	2.2	55	3.1
	40	3.4	45	2.1	50	2.6	55	3.5
	40	2.3	45	2.6	50	2.4	55	3.3
	40	5.0	45	3.8	51	2.0	57	3.5
	40	4.3	45	6.9	51	3.0	58	4.7

উপরের উপাত্তের ভিত্তিতে M. Rusticus এর ক্ষেত্রে

$$\bar{X}' = [38.80 \quad 2.68] \text{ এবং } S = \begin{bmatrix} 3.8933 & 0.5960 \\ 0.5960 & 1.2896 \end{bmatrix}$$

তাহলে

$$D = \begin{bmatrix} 1.9731 & 0 \\ 0 & 1.1356 \end{bmatrix}$$

এখন $Y_i = D^{-1}(X_i - \bar{X})$ অনুসারে

$$Y_1 = \begin{bmatrix} 0.5068 & 0 \\ 0 & 0.8806 \end{bmatrix} \left\{ \begin{pmatrix} 35 \\ 1.3 \end{pmatrix} - \begin{pmatrix} 38.80 \\ 2.68 \end{pmatrix} \right\} = \begin{bmatrix} -1.926 \\ -1.215 \end{bmatrix}$$

$$Y_2 = \begin{bmatrix} 0.5068 & 0 \\ 0 & 0.8806 \end{bmatrix} \left\{ \begin{pmatrix} 35 \\ 4.0 \end{pmatrix} - \begin{pmatrix} 38.80 \\ 2.68 \end{pmatrix} \right\} = \begin{bmatrix} -1.926 \\ 1.162 \end{bmatrix}$$

$$Y_3 = \begin{bmatrix} 0.5068 & 0 \\ 0 & 0.8806 \end{bmatrix} \left\{ \begin{pmatrix} 38 \\ 3.2 \end{pmatrix} - \begin{pmatrix} 38.80 \\ 2.68 \end{pmatrix} \right\} = \begin{bmatrix} -0.405 \\ 0.458 \end{bmatrix}$$

$$Y_4 = \begin{bmatrix} 0.5068 & 0 \\ 0 & 0.8806 \end{bmatrix} \left\{ \begin{pmatrix} 35 \\ 1.0 \end{pmatrix} - \begin{pmatrix} 38.80 \\ 2.68 \end{pmatrix} \right\} = \begin{bmatrix} -1.926 \\ -1.479 \end{bmatrix}$$

$$Y_5 = \begin{bmatrix} 0.5068 & 0 \\ 0 & 0.8806 \end{bmatrix} \left\{ \begin{pmatrix} 39 \\ 1.4 \end{pmatrix} - \begin{pmatrix} 38.80 \\ 2.68 \end{pmatrix} \right\} = \begin{bmatrix} 0.101 \\ -1.127 \end{bmatrix}$$

$$Y_6 = \begin{bmatrix} 0.5068 & 0 \\ 0 & 0.8806 \end{bmatrix} \left\{ \begin{pmatrix} 40 \\ 1.8 \end{pmatrix} - \begin{pmatrix} 38.80 \\ 2.68 \end{pmatrix} \right\} = \begin{bmatrix} 0.608 \\ -0.775 \end{bmatrix}$$

$$Y_7 = \begin{bmatrix} 0.5068 & 0 \\ 0 & 0.8806 \end{bmatrix} \left\{ \begin{pmatrix} 40 \\ 2.3 \end{pmatrix} - \begin{pmatrix} 38.80 \\ 2.68 \end{pmatrix} \right\} = \begin{bmatrix} 0.608 \\ -0.335 \end{bmatrix}$$

$$Y_8 = \begin{bmatrix} 0.5068 & 0 \\ 0 & 0.8806 \end{bmatrix} \left\{ \begin{pmatrix} 40 \\ 2.1 \end{pmatrix} - \begin{pmatrix} 38.80 \\ 2.68 \end{pmatrix} \right\} = \begin{bmatrix} 0.608 \\ -0.511 \end{bmatrix}$$

$$Y_9 = \begin{bmatrix} 0.5068 & 0 \\ 0 & 0.8806 \end{bmatrix} \left\{ \begin{pmatrix} 40 \\ 2.5 \end{pmatrix} - \begin{pmatrix} 38.80 \\ 2.68 \end{pmatrix} \right\} = \begin{bmatrix} 0.608 \\ -0.159 \end{bmatrix}$$

$$Y_{10} = \begin{bmatrix} 0.5068 & 0 \\ 0 & 0.8806 \end{bmatrix} \left\{ \begin{pmatrix} 40 \\ 3.5 \end{pmatrix} - \begin{pmatrix} 38.80 \\ 2.68 \end{pmatrix} \right\} = \begin{bmatrix} 0.608 \\ 0.772 \end{bmatrix}$$

$$Y_{11} = \begin{bmatrix} 0.5068 & 0 \\ 0 & 0.8806 \end{bmatrix} \left\{ \begin{pmatrix} 40 \\ 2.1 \end{pmatrix} - \begin{pmatrix} 38.80 \\ 2.68 \end{pmatrix} \right\} = \begin{bmatrix} 0.608 \\ -0.511 \end{bmatrix}$$

$$Y_{12}=\begin{bmatrix}0.5068 & 0\\ 0 & 0.8806\end{bmatrix}\left\{\begin{pmatrix}40\\3.4\end{pmatrix}-\begin{pmatrix}38.80\\2.68\end{pmatrix}\right\}=\begin{bmatrix}0.608\\0.634\end{bmatrix}$$

$$Y_{13}=\begin{bmatrix}0.5068 & 0\\ 0 & 0.8806\end{bmatrix}\left\{\begin{pmatrix}40\\2.3\end{pmatrix}-\begin{pmatrix}38.80\\2.68\end{pmatrix}\right\}=\begin{bmatrix}0.608\\-0.335\end{bmatrix}$$

$$Y_{14}=\begin{bmatrix}0.5068 & 0\\ 0 & 0.8806\end{bmatrix}\left\{\begin{pmatrix}40\\5.0\end{pmatrix}-\begin{pmatrix}38.80\\2.68\end{pmatrix}\right\}=\begin{bmatrix}0.608\\2.043\end{bmatrix}$$

$$Y_{15}=\begin{bmatrix}0.5068 & 0\\ 0 & 0.8806\end{bmatrix}\left\{\begin{pmatrix}40\\4.3\end{pmatrix}-\begin{pmatrix}38.80\\2.68\end{pmatrix}\right\}=\begin{bmatrix}0.608\\1.427\end{bmatrix}$$

এই পরিবর্তনকে অন্যভাবেও লেখা যায়। সেক্ষেত্রে

$$Y=[X-1\bar{X}']D^{-1}$$

$$=\left(\begin{bmatrix}35 & 1.3\\35 & 4.0\\38 & 3.2\\35 & 1.0\\39 & 1.4\\40 & 1.8\\40 & 2.3\\40 & 2.1\\40 & 2.5\\40 & 3.5\\40 & 2.1\\40 & 3.4\\40 & 2.3\\40 & 5.0\\40 & 4.3\end{bmatrix}-\begin{bmatrix}1\\1\\1\\1\\1\\1\\1\\1\\1\\1\\1\\1\\1\\1\\1\end{bmatrix}[38.80\ \ 2.68]\right)\begin{bmatrix}0.5068 & 0\\0 & 0.8806\end{bmatrix}$$

$$= \begin{bmatrix} -1.926 & -1.215 \\ -1.926 & 1.162 \\ -0.405 & 0.458 \\ -1.926 & -1.479 \\ 0.101 & -1.127 \\ 0.608 & -0.775 \\ 0.608 & -0.335 \\ 0.608 & -0.511 \\ 0.608 & -0.159 \\ 0.608 & 0.722 \\ 0.608 & -0.511 \\ 0.608 & 0.634 \\ 0.608 & -0.335 \\ 0.608 & 2.043 \\ 0.608 & 1.427 \end{bmatrix}$$

Mahalanobis পরিবর্তনের সূত্রকে লেখা যায়

$$Z = [\, X - 1\, \bar{X}' \,]\, S^{-1/2}$$

এখানে $S^{-1/2} = \Gamma \Lambda^{-1/2} \Gamma'$, $\Lambda^{-1/2} = \text{diag}(\lambda_j^{-1/2})$, λ_j হলো S ম্যাট্রিক্সের আইগেন মান এবং Γ হলো সমকৌণিক (orthogonal) ম্যাট্রিক্স যার স্তম্ভগুলো হলো আদর্শায়িত আইগেন ভেক্টর (standardized eigen vector) । আলোচিত S ম্যাট্রিক্স-এর আইগেন মান হলো $\lambda_1 = 4.02324$ এবং $\lambda_2 = 1.15966$ । λ_1 এর প্রাসঙ্গিক আইগেন ভেক্টর হলো $[\, 1 + 0.21802 \,]'$ এবং λ_2 এর ক্ষেত্রে তা $[\, 1 - 4.58664 \,]'$ ।

সুতরাং $$\Gamma = \begin{bmatrix} 0.97705 & 0.21302 \\ 0.21302 & -0.97705 \end{bmatrix}$$

এবং $$S^{-1/2} = \begin{bmatrix} 0.97705 & 0.21302 \\ 0.21302 & -0.97705 \end{bmatrix} \begin{bmatrix} 0.498554 & 0 \\ 0 & 0.928613 \end{bmatrix} \times \begin{bmatrix} 0.97705 & 0.21302 \\ 0.21302 & -0.97705 \end{bmatrix}$$

$$= \begin{bmatrix} 0.51807 & -0.08951 \\ -0.08951 & 0.90910 \end{bmatrix}$$

$$\therefore \quad Z = \begin{bmatrix} -3.80 & -1.38 \\ -3.80 & 1.32 \\ -0.80 & 0.52 \\ -3.80 & -1.68 \\ 0.20 & -1.28 \\ 1.20 & -0.88 \\ 1.20 & -0.38 \\ 1.20 & -0.58 \\ 1.20 & -0.18 \\ 1.20 & 0.82 \\ 1.20 & -0.58 \\ 1.20 & 0.72 \\ 1.20 & -0.38 \\ 1.20 & 2.32 \\ 1.20 & 1.62 \end{bmatrix} \times \begin{bmatrix} 0.51807 & -0.08951 \\ -0.08951 & 0.90910 \end{bmatrix}$$

$$= \begin{bmatrix} -1.845 & -0.914 \\ -2.089 & 1.540 \\ -0.461 & 0.550 \\ -1.818 & -1.187 \\ 0.218 & -1.182 \\ 0.700 & -0.907 \\ 0.656 & -0.453 \\ 0.674 & -0.635 \\ 0.638 & -0.271 \\ 0.548 & 0.638 \\ 0.674 & -0.635 \\ 0.557 & 0.547 \\ 0.656 & -0.453 \\ 0.075 & 2.002 \\ 0.477 & 1.365 \end{bmatrix}$$

প্রধান উপাদান পরিবর্তনের সূত্রকে লেখা যায়

$$W = \left[X - 1\bar{X} \right] \Gamma'$$

উপরে আলোচিত উপাত্তের ক্ষেত্রে

$$W = \begin{bmatrix} -4.0068 & 0.5388 \\ -3.4316 & -2.0992 \\ -0.6709 & -0.6785 \\ -4.0707 & 0.8320 \\ -0.0773 & 1.2932 \\ 0.9850 & 1.1158 \\ 1.0915 & 0.6269 \\ 1.0489 & 0.8223 \\ 1.1341 & 0.4315 \\ 1.3471 & -0.5456 \\ 1.0489 & 0.8223 \\ 1.3258 & -0.4479 \\ 1.0915 & 0.6269 \\ 1.6667 & -2.0111 \\ 1.5176 & -1.3272 \end{bmatrix}$$

এই W-এর স্তম্ভগুলোকে বলা হয় প্রধান উপাদান। উপাদানগুলো অসংশ্লেষিত এবং সেগুলো মূল চলকসমূহের রৈখিক সমাবেশ (linear combination)। প্রথম প্রধান উপাদানের ভেদাঙ্ক হলো $\lambda_1 = 4.02324$ এবং দ্বিতীয় প্রধান উপাদানের ভেদাঙ্ক হলো $\lambda_2 = 1.15966$। মূল উপাত্তের অধিকাংশ ভেদ সর্বোচ্চ ভেদাঙ্কবিশিষ্ট প্রধান উপাদানসমূহের মাধ্যমে ব্যাখ্যা করা যায়।

দ্বিতীয় অধ্যায়

বহুচলক বিন্যাস
(Multivariate Distribution)

২.১ দৈব ভেক্টর এবং এর ধর্ম (Random Vector and its Properties)

ধরা যাক x_{ij} হলো i-তম বস্তুর j-তম চলকের মান ($i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, p$ । আরো ধরা যাক $X_j = [x_{1j}\ x_{2j} \ldots x_{nj}]$ হলো j-তম চলকের মানভিত্তিক একটি ভেক্টর। এই X_j এর ভিত্তিতে একটি p-মাত্রার স্তম্ভ ভেক্টর হলো $X' = [X_1\ X_2 \ldots X_p]'$ । এই ভেক্টর X-এর বিন্যাস হলো বহুচলক বিন্যাস। এখানে ভেক্টরের কিছু গুণাবলি আলোচনা করা হবে।

ধরা যাক X_j হলো অবিচ্ছিন্ন দৈবচলক যার গড় হলো $E(X_j) = \mu_j$ এবং ভেদাঙ্ক হলো $E[X_j - \mu_j]^2 = \sigma_{jj}$ $(j = 1, 2, \ldots, p)$ । আরো ধরা যাক X_j ও X_l $(j \neq l = 1, 2, \ldots, p)$ এর সহ-ভেদাঙ্ক হলো $\mathrm{Cov}(X_j, X_l) = E[(X_j - \mu_j) \times (X_l - \mu_l)] = \sigma_{jl}$ । এখন ভেক্টর X এর প্রত্যাশিত মানকে লেখা যায়

$$E(X') = [E(X_1)\ E(X_2) \ldots E(X_p)]'$$

$$= [\mu_1\ \mu_2 \ldots \mu_p] = \mu'$$

$$\therefore \quad E(X) = \mu$$

এখানে μ হলো p-মাত্রার দৈব ভেক্টর X এর গড় ভেক্টর। ভেক্টর X এর সহ-ভেদাঙ্ক ম্যাট্রিক্স হলো

$$V(X) = E(X - \mu)(X - \mu)'$$

$$= E\begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_p - \mu_p \end{bmatrix} \begin{bmatrix} X_1 - \mu_1, & X_2 - \mu_2, & \ldots, & X_p - \mu_p \end{bmatrix}$$

$$= \begin{bmatrix} E(X_1 - \mu_1)^2 & E(X_1 - \mu_1)(X_2 - \mu_2) & \ldots & E(X_1 - \mu_1)(X_p - \mu_p) \\ E(X_1 - \mu_1)(X_2 - \mu_2) & E(X_2 - \mu_2)^2 & \ldots & E(X_2 - \mu_2)(X_p - \mu_p) \\ \ldots & \ldots & \ldots & \ldots \\ E(X_1 - \mu_1)(X_p - \mu_p) & E(X_2 - \mu_2)(X_p - \mu_p & \ldots & E(X_p - \mu_p)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \cdots & \sigma_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \cdots & \sigma_{pp} \end{bmatrix} - \sum$$

আবার ধরা যাক P-মাত্রার দৈব চলক X হলো একটি ভেক্টর, যেখানে X-কে লেখা যায়

$$X' = (X_1, X_2, \ldots, X_p)$$

ভেক্টর X এর i-তম উপাদান $X_i (i = 1, 2, \ldots, p)$ হলো যে কোনো বস্তুর i-তম বৈশিষ্ট্য। যে কোনো বস্তুর সকল (p) বৈশিষ্ট্য যুগপৎ পরিমাপ করা হলে p-মাত্রার দৈব চলক ভেক্টরের উদ্ভব হয়। এই চলকসমূহের ভিত্তিতে কোনো দৈব ঘটনা (random event) সংজ্ঞায়িত করা হলে ঐ ঘটনার সম্ভাবনা বা ঘটনাসমূহের সম্ভাবনা ক্রমসঞ্চিত বিন্যাস ফাংশন (cumulative distribution function) হতে নির্ণয় করা যায়। এখানে

$$F(x_1, x_2, \ldots, x_p) = p[X_1 \leqslant x_1, X_2 \leqslant x_2, \ldots, X_p \leqslant X_p]$$

হলো সকল বাস্তব সংখ্যা [real number] $x_1, x_2, \ldots, x_p$ এর গুচ্ছের ক্রম-সঞ্চিত বিন্যাস ফাংশন। এই বিন্যাস ফাংশন $F(x_1, x_2, \ldots, x_p)$ অনপেক্ষভাবে অবিচ্ছিন্ন (absolutely continuous) হলে ঘনত্ব ফাংশন (density function) এর সংজ্ঞা হলো

$$f(x_1, x_2, \ldots, x_p) = \frac{\partial^p F(x_1, x_2, \ldots, x_p)}{\partial x_1 \, \partial x_2 \ldots \partial x_p}$$

কিন্তু X এর উপাদানসমূহ বিচ্ছিন্ন চলক হলে

$$f(x_1, x_2, \ldots, x_p) = p\{X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p\}$$

অর্থাৎ X বিচ্ছিন্ন চলকের ভেক্টর হলে $X = (x_1, x_2, \ldots, x_p)$ এর ঘনত্ব ফাংশন হবে $P(X = x)$।

অনেক সময় পুরো গুচ্ছ $\{X_1, X_2, \ldots, X_p\}$ এর পরিবর্তে তার উপ-গুচ্ছ $\{X_1, X_2, \ldots, X_r\}$ $(r < p)$ এর ঘনত্ব ফাংশন নির্ণয় করার প্রয়োজন হয়। ধরা যাক $F(X_1, X_2, \ldots, X_p)$ হলো $\{X_1, X_2, \ldots, X_p\}$ এর ক্রমসঞ্চিত ঘনত্ব ফাংশন। তাহলে $\{X_1, X_2, \ldots, X_r\}$ এর প্রান্তিক ক্রমসঞ্চিত ঘনত্ব ফাংশন হবে

$$\begin{aligned} P &= [X_1 \leqslant x_1, X_2 \leqslant x_2, \ldots, X_r \leqslant x_r] \\ &= P[X_1 \leqslant x_1, X_2 \leqslant x_2, \ldots, X_r \leqslant x_r, X_{r+1} \leqslant \infty, X_p \leqslant \infty] \\ &= F[x_1, x_2, \ldots, x_r, \infty, \ldots, \infty] \\ &= F(x_1, x_2, \ldots, x_r) \end{aligned}$$

এক্ষেত্রে $\{X_1, X_2, \ldots, X_r\}$ এর প্রান্তিক ঘনত্ব ফাংশন হলো

$$f(x_1, x_2, \ldots, x_r) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(t_1, t_2, \ldots, t_p) dt_{r+1} \ldots dt_p$$

কিন্তু $\{X_1, X_2, \ldots, X_p\}$ বিচ্ছিন্ন চলক গুচ্ছ হলে

$$f(x_1, x_2, \ldots, x_r) = \sum_{S_{r_1}} \sum_{S_{r_2}} \ldots \sum_{S_{r_p}} f(x_1, x_2, \ldots, x_p)$$

এখানে S_{r_i} হলো X_1 এর সম্ভাব্য সকল মানের গুচ্ছ।

অনপেক্ষতা (Independence) : চলক গুচ্ছ $\{X_1, X_2, \ldots, X_p\}$ পরস্পর অনপেক্ষ হবে যদি তাদের ক্রমসঞ্চিত ঘনত্ব ফাংশন নিম্নরূপ হয় :

$$F(x_1, x_2, \ldots, x_p) = F_1(x_1)\, F_2(x_2) \ldots, F_p(x_p)$$

এখানে $F_i(.)$ হলো X_i এর প্রান্তিক ক্রমসঞ্চিত ঘনত্ব ফাংশন। আবার,

$$f(x_1, x_2, \ldots, x_p) = \frac{\partial^p F(x_1, x_2, \ldots, x_p)}{\partial x_1\, \partial x_{2r} \ldots \partial x_p}$$

কিন্তু $X_1, X_2, \ldots, X_p$ পরস্পর অনপেক্ষ হলে

$$f(x_1, x_2, \ldots, x_p) = \frac{\partial^p F_1(x_1)\, F_2(x_2) \ldots F_p(x_p)}{\partial x_1\, \partial x_2 \ldots \partial x_p}$$

$$= \frac{dF_1(x_1)\, d\, F_2(x_2) \ldots d\, F_p(x_p)}{dx_1\, dx_2 \ldots dx_p}$$

$$= f_1(x_1)\, f_2(x_2) \ldots f_p(x_p)$$

চলকসমূহ $X_1, X_2, \ldots, X_p$ অনপেক্ষ হলে অনপেক্ষতার সূত্র প্রয়োগ করে ঐ চলকসমূহের যে কোনো ব্যাপ্তিভিত্তিক বহুচলক সম্ভাবনা (multivariate probabilities) নিম্নরূপভাবে নির্ণয় করা যায় :

ধরা যাক $R_1, R_2, \ldots, R_p$ হলো P ব্যাপ্তি। মনে করি এই ব্যাপ্তিভিত্তিক যুগ্ম ঘটনা হলো $X_i \in R_i$; $i = 1, 2, \ldots, p$। এখন $X_1, X_2, \ldots, X_p$ অবিচ্ছিন্ন চলক হলে

$$P(X_1 \in R_1, X_2 \in R_2, \ldots, X_p \in R_p)$$

$$= \int_{R_1} \int_{R_2} \ldots \int_{R_p} f(t_1, t_2, \ldots, t_p)\, dt_1\, dt_2 \ldots dt_p$$

$$= \int_{R_1} f_1(t_1)dt_1 \int_{R_2} f_2(t_2)dt_2 \ \dots \ \int_{R_p} f(t_p)dt_p$$

$$= P(X_1 \in R_1)\ P(X_2 \in R_2)\ \dots\ P(X_p \in R_p)$$

শর্তাধীন ঘনত্ব ফাংশন (Conditional Density Function) : ধরা যাক $Z = (X_1, X_2)$ এর যুগ্ম ঘনত্ব ফাংশন হলো $f(x_1, x_2)$ এবং X_1 ও X_2 এর প্রান্তিক ঘনত্ব ফাংশন হলো যথাক্রমে $f_1(x_1)$ এবং $f_2(x_2)$। ধরা যাক X_1 এর মান (x_{11}, x_{12}) ব্যাপ্তির মধ্যে হলে একটি ঘটনা A এর সংজ্ঞায়ন করা যায় এবং X_2 এর মান (x_{21}, x_{22}) ব্যাপ্তির মধ্যে হলে অন্য একটি ঘটনা B সংজ্ঞায়ন করা যায়। তাহলে, শর্তাধীন সম্ভাবনার সূত্র প্রয়োগ করে লেখা যায়

$$P(A/B) = \frac{P\{x_{11} \leqslant X_1 \leqslant x_{12} \mid x_{21} \leqslant X_2 \leqslant x_{22}\}}{}$$

$$= \frac{\int_{x_{11}}^{x_{12}} \int_{x_{21}}^{x_{22}} f(s, t)ds\, dt}{\int_{x_{21}}^{x_{22}} f_2(t)dt}$$

এখন Anderson (1958) এর কাজ অনুসরণ করে লেখা যায়

$$\int_{x}^{x+\Delta x} f_2(t)dt = f_2(x^*)\,\Delta x$$

এখানে $x_{21} = x$, $x_{22} = x + \Delta x$ বিবেচনা করা হয়েছে এবং চলক X_2-কে অবিচ্ছিন্ন বিবেচনা করা হয়েছে, $x \leqslant x^* \leqslant x + \Delta x$। অনুরূপভাবে লেখা যায়

$$\int_{x}^{x+\Delta x} f(s, t)\, ds\, dt = f(s, x(s))\,\Delta x$$

এখানে $x \leqslant x(s) \leqslant x + \Delta x$। সুতরাং

$$P\{x_{11} < X_1 \leqslant x_{12} \mid x \leqslant X_2 \leqslant x + \Delta x\} = \frac{\int_{x_{11}}^{x_{12}} f(s, x(s))}{f_2(x^*)} ds$$

এই সূত্রের জন্য শর্ত হলো $f_2(x) > 0$। এখন $\lim \Delta x \to 0$ হলে লেখা যায়

$$P\{x_{11} < X \leqslant x_{12} \mid X_2 = x\} = \int_{x_{11}}^{x_{12}} f(s/x)ds$$

এখানে $f(s/x) = f(s, x)/f_2(x)$ । সুতরাং, যে কোনো দেয়া মান x এর জন্য $f(s/x)$ হলো $X_2 = x$ শর্তাধীনে X_1 এর শর্তাধীন ঘনত্ব ফাংশন। কিন্তু X_1 ও X_2 অনপেক্ষ হলে

$$f(x_1/x_2) = f_1(x_1)$$

উপরে আলোচিত শর্তাধীন ঘনত্ব ফাংশন-এর সূত্র বহুচলকের ক্ষেত্রেও প্রয়োগ করা যায়। ধরা যাক

$X = \{X_1, X_2, \ldots, X_p\} = \{X_1, X_2, \ldots, X_r, X_{r+1} X_{r+2}, \ldots, X_p\}$

এখন $X_{r+1} = x_{r+1}, X_{r+2} = x_{r+2}, \ldots, X_p = x_p$ দেয়া থাকলে $X_1, X_2, \ldots, X_r$ এর শর্তাধীন ঘনত্ব ফাংশন হবে

$$\frac{f(x_1, x_2, \ldots, x_p)}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} f(s_1, s_2, \ldots, s_r, s_{r+1}, \ldots, s_p)ds_1 \ldots ds_p}$$

নিয়ামক ফাংশন (Characteristic Function) : ধরা যাক X হলো একটি P-মাত্রার দৈব ভেক্টর। এর নিয়ামক ফাংশন $\varphi(t)$ এর সংজ্ঞা হলো

$$\varphi(t) = E\left[e^{it'X}\right] = \int e^{it'X} f(X)dX$$

এখানে t হলো বাস্তব মান $t_1, t_2 \ldots, t_p$ ভিত্তিক একটি স্তম্ভ ভেক্টর। এই নিয়ামক ফাংশন সব সময় নির্ণয় করা যায়, অর্থাৎ $\varphi(0) = 1$ এবং $|\varphi(t)| \leqslant 1$. নিয়ামক ফাংশনের কুমুল্যান্ট উৎপাদনকারী ফাংশন হলো $\log_e \varphi(t)$ এবং একে বিযোজন করে X_j $(j = 1, 2, \ldots, p)$ এর পরিঘাত নির্ণয় করা যায়। যেমন

$$E(X_j, X_k) = -\left.\frac{\partial^2 \log_e \varphi(t)}{\partial t_j t_k}\right|_{t=0}, \quad j \neq k$$

তাছাড়া inversion উপপাদ্য প্রয়োগ করে নিয়ামক ফাংশন হতে ঘনত্ব ফাংশন নির্ণয় করা যায়। ধরা যাক X হলো দৈব ভেক্টর যার ঘনত্ব ফাংশন হলো $f(X)$ এবং নিয়ামক ফাংশন হলো $\varphi(t)$। তাহলে

$$f(X) = \frac{1}{(2\pi)^p}\int_{-\infty}^{\infty} e^{it'X}\varphi(t)dt$$

নিয়ামক ফাংশন এর আরো একটি বৈশিষ্ট্য হলো যে, দৈব ভেক্টর X-কে দুটি উপ-ভেক্টরে প্রকাশ করা সম্ভব হলে অর্থাৎ $X = (X_1', X_2')$ হলে

$$\varphi(t) = \varphi_1(t_1)\varphi_2(t_2)$$

হবে যদি এবং কেবল যদি X_1 ও X_2 অনপেক্ষ হয়। এখানে $\varphi_1(t_1)$ এবং $\varphi_2(t_2)$ হলো X_1 ও X_2 এর নিয়ামক ফাংশন, যথাক্রমে এবং $t' = (t_1', t_2')$।

আবার X_1 ও X_2 দুটি p-মাত্রার অনপেক্ষ ভেক্টর হলে এবং $\varphi_1(t)$ ও $\varphi_2(t)$ যথাক্রমে X_1 ও X_2 এর নিয়ামক ফাংশন হলে, $X_1 + X_2$ এর নিয়ামক ফাংশন হবে

$$\varphi_{X_1+X_2}(t) = \varphi_1(t)\,\varphi_2(t)$$

দুটি ভেক্টর X_1 ও X_2 এর নিয়ামক ফাংশন $\varphi_1(t)$ ও $\varphi_2(t)$ একই হবে, যদি X_1 ও X_2 একই বিন্যাস অনুসরণ করে।

চলক পরিবর্তন (Transformation of Variable) : ধরা যাক ভেক্টর $X' = (X_1, X_2, \ldots, X_p)$ এর ক্ষেত্রে $X_1, X_2, \ldots, X_p$ এর যুগ্ম বিন্যাস হলো $f(x_1, x_2, \ldots, x_p)$। এখন বাস্তব মানভিত্তিক একটি ফাংশন

$$y_j = y_j(x_1, x_2, \ldots, x_p) \; ; \; j = 1, 2, \ldots, p$$

বিবেচনা করা যাক। এখানে বিবেচনা করা যাক যে, x থেকে y-এ পরিবর্তন ওয়ান-টু-ওয়ান। সুতরাং বিপরীত পরিবর্তনের মাধ্যমে পাওয়া যায়

$$x_1 = x_1(y_1, y_2, \ldots, y_p)$$

এখন দৈবচলক $Y_1, Y_2, \ldots, Y_p$ যদি $X_1, X_2, \ldots, X_p$ এর ফাংশন

$$Y_1 = y_1(X_1, X_2, \ldots, X_p)$$

হয়, তাহলে $Y_1, Y_2, \ldots, Y_p$ এর যৌথ ফাংশন হবে

$$g(y_1, y_2, \ldots, y_p) = f[x_1(y_1, y_2, \ldots, y_p), x_2(y_1, y_2, \ldots, y_p), \ldots x_p(y_1, y_2, \ldots, y_p)]\, J(y_1, y_2, \ldots, y_p)$$

এখানে J হলো পরিবর্তনের Jacobian এবং এর মান হলো

$$J(y_1, y_2, \ldots, y_p) = \text{মড্} \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_p}{\partial y_p} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial x_p}{\partial y_1} & \frac{\partial x_p}{\partial y_2} & \cdots & \frac{\partial x_p}{\partial y_p} \end{vmatrix}$$

ভেক্টরের আরো কিছু ধর্ম (More Properties of Vectors)

দৈব ভেক্টরের গুণ (Product of a random vector) : ধরা যাক C হলো একটি ধ্রুবক এবং X হলো একটি ভেক্টর। তাহলে, CX হলো ভেক্টরের স্কেলার গুণ।

দুটি ভেক্টরের স্কেলার গুণ (Scalar product of two vectors) : ধরা যাক $X=(X_1, X_2, \dots, X_p)'$ এবং $Y=(Y_1, Y_2, \dots, Y_p)'$ দুটি p-মাত্রার ভেক্টর। তাদের স্কেলার গুণ হবে

$$X'Y=X_1Y_1+X_2Y_2+\dots+X_pY_p=\sum_{j=1}^{p}X_jY_j$$

যদি A একটি $n\times p$ ম্যাট্রিক্স হয় এবং X একটি p-মাত্রার ভেক্টর হয়, তাহলে AX-কে বলা হয় A দ্বারা X-এর গুণ। কিন্তু A যদি ধ্রুবক হয় এবং $E(X)=\mu$ হলে $E(AX)=A\mu$ হবে এবং $V(AX)=A^2V(X)=A^2\Sigma$। আবার A যদি $a_1, a_2, \dots, a_p$ ধ্রুবকসমূহবিশিষ্ট একটি স্তম্ভ ভেক্টর হয়, তাহলে ভেক্টর A ও ভেক্টর X এর গুণ হলো

$$A'X=\sum_{j=1}^{p}a_jX_j$$

এবং
$$E(A'X)=A'E(X)=A'\mu=\Sigma a_j\mu_j$$

$$V(A'X)=V(\Sigma a_jX_j)=\Sigma a_j^2\,V(X_j)+\sum_{j\neq k}\sum a_ja_k\,Cov(X_jX_k)$$

$$=\Sigma a_j^2\sigma_{jj}+\sum_{j\neq k}\sum a_ja_k\sigma_{jk}$$

$$=A'\Sigma A$$

অনুরূপভাবে একটি B ভেক্টর বিবেচনা করা হলে, যেখানে $B=(b_1, b_2, \dots, b_p)'$, তাহলে B ও X এর গুণ হলো BX। সে ক্ষেত্রে

$$Cov(A'X, B'X)=Cov\left(\sum a_jX_j, \sum b_kX_k\right)$$

$$=\sum_{j}^{p}\sum_{k}^{p}a_jb_k\,\sigma_{jk}$$

$$=A'\Sigma B$$

কিন্তু X-কে ম্যাট্রিক্স A দ্বারা গুণ করা হলে

$$E(AX) = AE(X) = A\mu$$

এবং

$$V(AX) = E(AX - A\mu)\,(AX - A\mu)'$$
$$= AE(X-\mu)\,(X-\mu)'A'$$
$$= A\Sigma A'$$

Kronecker গুণ (Kronecker Product) : ধরা যাক $A = (a_{ij})$ এবং $B = (b_{kl})$ হলো, যথাক্রমে $(m \times n)$ এবং $(p \times q)$ অর্ডারের দুটি ম্যাট্রিক্স। তাহলে A ও B এর Kronecker গুণ হলো $A \times B$ যা লেখা যায়

$$\begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \dots & \dots & \dots & \dots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{bmatrix}$$

এই ম্যাট্রিক্স-এর অর্ডার হলো $(mp \times nq)$ ।

এখন উপাত্ত ম্যাট্রিক্স X-কে এমনভাবে একটি স্তম্ভে সাজানো যাক, যেখানে সাজানো ভেক্টর হবে X ম্যাট্রিক্স-এর স্তম্ভগুলোর স্তূপাকার। অর্থাৎ

$$X^V = \begin{bmatrix} X_{(1)} \\ X_{(2)} \\ \vdots \\ X_{(p)} \end{bmatrix}$$

তাহলে, $V(X^V) = \Sigma \times$ হলো $(np \times np)$ অর্ডারের সহ-ভেদাঙ্ক ম্যাট্রিক্স। এখন Kronecker গুণ-এর ধর্ম হতে লেখা যায়

(i) যে কোনো ধ্রুবক α-এর ক্ষেত্রে

$$\alpha(A \times B) = (\alpha A) \times B = A \times \alpha B = \alpha A \times B$$

(ii) $A \times B \times C = (A \times B) \times C = A \times B \times C$

(iii) $(A \times B)' = A' \times B'$

(iv) $(A \times B)\,(F \times G) = (AF) \times (BG)$

(v) $(A \times B)^{-1} = A^{-1} \times B^{-1}$ । এখানে A ও B-কে নন্-সিঙ্গুলার হতে হবে।

(vi) $(A+B) \times C = A \times C + B \times C$

(vii) $A \times (B+C) = A \times B + A \times C$

(viii) $(A \times B)^V = (B' \times A) X^V$

এখন $E(X) = \mu$ হলে

$$E[(A \times B)^V] = (B' \times A)\, \mu \times I = B'\mu \times AI$$

$$V[(A \times B)^V] = (B' \times A)\,(\Sigma \times I)\,(B' \times A)'$$

$$= B'\Sigma B \times AA'$$

Mahalanobis দূরত্ব (Mahalanobis Distance) : ধরা যাক $X_1 \sim (\mu_1, \Sigma)$ এবং $X_2 \sim (\mu_2, \Sigma)$ হলো দুটি ভেক্টর যাদের গড় ভেক্টর হলো যথাক্রমে μ_1 ও μ_2 এবং সাধারণ সহ-ভেদাঙ্ক ম্যাট্রিক্স Σ । এক্ষেত্রে μ_1 ও μ_2 এর মধ্যে Mahalanobis দূরত্ব হলো

$$\Delta^2(\mu_1, \mu_2) = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) \qquad (২.১.১)$$

এর বর্গমূল ।

এই দূরত্ব নিম্নরূপ পরিবর্তনের ফলে পরিবর্তিত হয় না ।

$$X \to AX + b,\ Y \to AY + b,\ \Sigma \to A\Sigma A'$$

এখানে A হলো নন্-সিঙ্গুলার ম্যাট্রিক্স ।

২.২ বহুচলক পরিমিত বিন্যাস (Multivariate Normal Distribution)

ধরা যাক, X হলো p-মাত্রার একটি vector । তাহলে X ভেক্টরের যুগ্ম সম্ভাবনা ঘনত্ব ফাংশন (p.d.f) হলো

$$f(x_1, x_2, \ldots, x_p) = |2\pi\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)}$$

$$= \frac{[\,|R|\,]^{1/2}}{(2\pi)^{p/2}} e^{-\frac{1}{2}(X-\mu)'R(X-\mu)},$$

$$-\infty < x_i < \infty \qquad (২.২.১)$$

এখানে $\Sigma > 0$, $R = \Sigma^{-1}$, μ হলো X ভেক্টরের গড় ভেক্টর । $R = (r_{ij})$ হলো ধনাত্মক ডেফিনিট (positive definite) এবং r_{ij} গুলো হলো ধ্রুবক ।

যেহেতু $|R|$ ধনাত্মক, $|R|^{1/2}/(2\pi)^{p/2}$ ধনাত্মক এবং e-এর যে কোনো ঘাত-এর মান ধনাত্মক হওয়াতে (২.২.১) ধনাত্মক । অর্থাৎ $f(x_1, x_2, \ldots, x_p) > 0$ । এখন (২.২.১) একটি ঘনত্ব ফাংশন হবে যদি

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_p)\, dx_1\, dx_2 \ldots dp = 1$$

হয়। এটি প্রমাণ করার জন্য একটি নতুন চলক ভেক্টর Z ধরা যাক, যেখানে $Z = X - \mu$। এটি হতে লেখা যায় $Z_j = x_j - \mu_j$; $j = 1, 2, \ldots, p$। কিন্তু $dx_j/dz_k = 0$ $(j \neq k)$ এবং $dx_j/dz_j = 1$ হওয়ার কারণে X চলক হতে Z চলকে পরিবর্তনের Jacobian হবে এক (unity)। সুতরাং,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{|R|^{1/2}}{(2\pi)^{p/2}} e^{-\frac{1}{2}(X-\mu)'R(X-\mu)}\, dx_1\, dx_2 \cdots dx_p$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{|R|^{1/2}}{(2\pi)^{p/2}} e^{-\frac{1}{2}Z'RZ}\, dz_1\, dz_2 \cdots dz_p$$

এখানে X থেকে Z-এ পরিবর্তন ওয়ান্-টু-ওয়ান্ এবং সমপূর্ণ X-স্পেস-এ পরিবর্তিত হয়। এখন Z-এর উপর একটি সমকৌণিক পরিবর্তন (orthogonal transformation) $Z = PT$ করা যাক, এখানে P হলো সমকৌণিক ম্যাট্রিক্স এবং $T = \{t_j\}$ হলো একটি ভেক্টর। তাহলে

$$Z'RZ = T'P'RPT = T'DT$$

কারণ, R ধনাত্মক ডেফিনিট হওয়াতে $P'RP = D$, এখানে D হলো R ম্যাট্রিক্স-এর নিয়ামক মূল (characteristic root) বিশিষ্ট কৌণিক (diagonal) ম্যাট্রিক্স। তাহলে, পাওয়া যায়

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{|R|^{1/2}}{(2\pi)^{p/2}} e^{-\frac{1}{2}Z'RZ}\, dz_1\, dz_2 \cdots dz_p$$

$$= \frac{|R|^{1/2}}{(2\pi)^{p/2}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{-\frac{1}{2}T'DT}\, dt_1\, dt_2 \cdots dtp$$

$$= \frac{|R|^{1/2}}{(2\pi)^{p/2}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}d_1t_1^2}\, dt_1 \int_{-\infty}^{\infty} e^{-\frac{1}{2}d_2t_2^2}\, dt_2 \cdots$$

$$\cdots\int_{-\infty}^{\infty} e^{-\frac{1}{2}d_pt_p^2}\, dtp \quad -\infty < t_j < \infty$$

কিন্তু এখানে $D = \text{diag}(d_j)$। কিন্তু এক চলক পরিমিত বিন্যাস থেকে লেখা যায়

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}d_j t_j^2}\, dt_j = \sqrt{\frac{2\pi}{d_j}}$$

$$\therefore \quad \frac{|R|^{1/2}}{(2\pi)^{p/2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}d_1 t_1^2} dt_1 \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}d_p t_p^2} dt_p$$

$$= \frac{|R|^{1/2}}{(2\pi)^{p/2}} \cdot \frac{(2\pi)^{p/2}}{\prod_{j=1}^{p} (d_j)^{1/2}}$$

আবার

$$|D| = \prod_{j=1}^{p} d_j, \quad |D|^{1/2} = \prod_{j=1}^{p} (d_j)^{1/2}, \quad |R| = |P'RP| = |D|$$

সুতরাং $\quad |R|^{1/2} = \prod_{j=1}^{p} d_j^{1/2}$ । কাজেই

$$\frac{|R|^{1/2}}{(2\pi)^{p/2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}d_1 t_1^2} dt_1 \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}d_p t_p^2} d_p = 1$$

সুতরাং, ২.২.১ হলো একটি ঘনত্ব ফাংশন।

দৈব চলকসমূহের গড় (Means of the Random Variables) : লক্ষ্য করা গিয়েছে যে,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{|R|^{1/2}}{(2\pi)^{p/2}} e^{-\frac{1}{2}(X-\mu)\Sigma^{-1}(X-\mu)} dx_1, dx_2 \cdots dx_p = 1$$

অথবা

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\left[|\sigma^{jk}|\right]^{1/2}}{(2\pi)^{p/2}} e^{-\frac{1}{2}\sum_{j,k}^{p} \sigma^{jk}(x_j-\mu_j)(x_k-\mu_k)} \times dx_1\, dx_2 \cdots dx_p = 1 \qquad (২.২.২)$$

এখন μ_j এর প্রসঙ্গে উভয় পাশে বিয়োজন করে পাওয়া যায়

$$E\left[\sum_{j}^{p} \sigma^{jk}(x_j-\mu_j)\right] = 0 ; \quad j = 1, 2, \cdots, p$$

অথবা $$\sum_{j} \sigma^{jk} E(x_j - \mu_j) = 0 \qquad (২.২.৩)$$

কিন্তু R ধনাত্মক ডেফিনিট হওয়াতে $| R | = | \sigma^{jk} | \neq 0$ । সুতরাং, ২.২.৩ এর সমাধান হবে

$$E(x_j) = \mu_j \; ; \quad j = 1, 2, \cdots, p$$

অর্থাৎ $\mu_1, \mu_2, \cdots, \mu_p$ হলো যথাক্রমে $X_1, X_2, \cdots, X_p$ এর গড় ।

দৈব চলকসমূহের সহ-ভেদাঙ্ক (Covariances of Random Variables) : ঘনত্ব ফাংশনের সংজ্ঞা হতে পাওয়া যায়

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\sum_{j}^{p} \sigma^{jk}(x_j - \mu_j)(x_k - \mu_k)} dx_1 \cdots dx_p = \frac{(2\pi)^{p/2}}{[|\sigma^{jk}|]^{1/2}}$$

এখন উভয় পাশে σ^{jk} এর প্রসঙ্গে বিযোজন করে এবং

$$-(1 + \delta_{jk}) \frac{[|\sigma^{jk}|]^{1/2}}{(2\pi)^{p/2}}$$

দ্বারা গুণ করে পাওয়া যায়

$$E[(x_j - \mu_j)(x_k - \mu_k)] = \sigma_{jk}$$

এটি $j = k$ এবং $j \neq k$ উভয় ক্ষেত্রেই সত্য । এখানে δ_{jk} হলো Kronecker δ । সুতরাং, x_j ও x_k এর সহ-ভেদাঙ্ক হলো σ_{jk} ।

প্রান্তিক বিন্যাসসমূহ (Marginal Distributions) : প্রথমে x_1 এর বিন্যাস নির্ণয় করা যাক । এই x_1 এর p.d.f. হবে

$$f_1(x_1) = \frac{|R|^{1/2}}{(2\pi)^{p/2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}(X-\mu)' R(X-\mu)} \times dx_2\, dx_3 \cdots dx_p$$

এখানে দ্বিপদিক ফাংশন (Quadratic Function)

$(X - \mu)' R(X - \mu)$-কে লেখা যায়

$$(X-\mu)' R(X-\mu) = [(x_1 - \mu_1), X_2 - M_2)'] \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ X_2 - M_2 \end{bmatrix}$$

$$= (x_1 - \mu_1) r_{11} (x_1 - \mu_1) + (x_1 - \mu_1) R_{12} (X_2 - M_2)$$

$$+ (X_2 - M_2)' R_{21} (x_1 - \mu_1) + (X_2 - M_2)' R_{22} (X_2 - M_2)$$

এখানে X_2 ও M_2 হলো যথাক্রমে X ও μ ভেক্টরের শেষ $(p-1)$ মানবিশিষ্ট ভেক্টর, R_{11} হলো R ম্যাট্রিক্স-এর প্রথম সারির প্রথম স্তম্ভের মান। কিন্তু R প্রতিসম (symmetric) এবং ধনাত্মক ডেফিনিট হওয়াতে $R_{12}=R_{21}'$ এবং R_{22}^{-1} পাওয়া যায়। এই R_{22}^{-1} প্রতিসম। আবার, $R_{11}=r_{11}>0$। তাহলে, লেখা যায়

$$(X-\mu)' R(X-\mu)=(x_1-\mu_1)(R_{11}-R_{12}R_{22}^{-1}R_{21})(x_1-\mu_1)$$
$$+\{[(X_2-M_2)+R_{22}^{-1}R_{21}(x_1-\mu_1)]'$$
$$\times R_{22}[(X_2-M_2)+R_{22}^{-1}R_{21}(x_1-\mu_1)]\}$$

সুতরাং, $f_1(x_1)$-কে লেখা যায়

$$f_1(x_1)=\frac{|R|^{1/2}}{(2\pi)^{p/2}}K\,e^{-\frac{1}{2}(x_1-\mu_1)(R_{11}-R_{12}R_{22}^{-1}R_{21})(x_1-\mu_1)}$$

এখানে

$$K=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2}\left\{X_2-\left[M_2-R_{22}^{-1}R_{21}(x_1-\mu_1)\right]\right\}'\right.$$
$$\left.\times R_{22}\{X_2-[M_2-R_{22}^{-1}R_{21}(x_1-\mu_1)]\}dX_2\right.$$

$dX_2=dx_2\,dx_3\ldots dxp$। এখানে x_1 সমাকলনযোগ্য নয়। একে সমাকলনের প্রসঙ্গে ধ্রুবক বিবেচনা করা যায়। এখন, ধরা যাক

$$M_2-R_{22}^{-1}R_{21}(x_1-\mu_1)=H,\ \text{তাহলে}$$

$$f_1(x_1)=\frac{|R|^{1/2}}{(2\pi)^{p/2}}K\,e^{-\frac{1}{2}(x_1-\mu_1)(R_{11}-R_{12}R_{22}^{-1}R_{21})(x_1-\mu_1)}$$

এখানে $K=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}e^{-\frac{1}{2}(X_2-H)'R_{22}(X_2-H)}\,dX_2$

$$=\frac{(2\pi)^{\frac{1}{2}(p-1)}}{|R_{22}|^{1/2}}$$

$$\therefore\quad f_1(x_1)=\frac{|R|^{1/2}}{(2\pi)^{p/2}}\,\frac{(2\pi)^{\frac{1}{2}(p-1)}}{|R_{22}|^{p/2}}$$
$$\times e^{-\frac{1}{2}(x_1-\mu_1)(R_{11}-R_{12}R_{22}^{-1}R_{21})(x_1-\mu_1)}$$

কিন্তু $|R| = |R_{22}| \cdot |R_{11} - R_{12} R_{22}^{-1} R_{21}|$। এখানে $(x_1 - \mu_1)$ এবং $|R_{11} - R_{12} R_{22}^{-1} R_{21}|$ হলো স্কেলার (scalar)। ধরা যাক $(R_{11} - R_{12} R_{22}^{-1} R_{21}) = 1/\sigma_1^2$। এটি ধনাত্মক ডেফিনিট। কাজেই

$$f_1(x_1) = \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{1}{2}\sigma_1^2 (x_1 - \mu_1)^2}, \quad -\infty < x_1 < \infty$$

এটি একচলক পরিমিত বিন্যাসের ঘনত্ব ফাংশন। অনুরূপভাবে যে কোনো $x_j (j = 1, 2, \ldots, p)$ এর p.d.f. নির্ণয় করা যেতে পারে।

উপরে x_1 এর বিন্যাস নির্ণয় করার জন্য X ভেক্টরকে x_1 ও X_2 দুটি ভাগে বিভক্ত করা হয়েছে। এখন x_1 এর পরিবর্তে X_1 নামক কোন উপ-ভেক্টর (sub-vector)-এর p.d.f. নির্ণয় করতে হলে X ভেক্টরকে X_1 ও X_2 দুটি উপ-ভেক্টরে ভাগ করা যেতে পারে। সেক্ষেত্রে দ্বিপদিক ফাংশন $(X-\mu)' R(X-\mu)$-কে লেখা যায়

$$(X-\mu)' R(X-\mu) = [(X_1 - M_1), (X_2 - M_2)]' \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \begin{bmatrix} X_1 - M_1 \\ X_2 - M_2 \end{bmatrix}$$

ধরা যাক X_1 ভেক্টরে $s (< p)$ মান আছে এবং $(X_1 - \mu_1)$ হলো s-মাত্রার ভেক্টর, R_{11} হলো $(s \times s)$ অর্ডারের ম্যাট্রিক্স। তাহলে, X_1 এর p.d.f. হবে

$$g(X_1) = \frac{\exp[-\frac{1}{2}(X_1 - M_1)' (R_{11} - R_{12} R_{22}^{-1} R_{21})(X_1 - M_1)]}{(2\pi)^{s/2} |R_{11} - R_{12} R_{22}^{-1} R_{21}|^{-1/2}}$$

$$-\infty < x_j < \infty; \quad j = 1, 2, \ldots, s$$

শর্তাধীন বিন্যাস (Conditional Distribution) : ভেক্টর X-এর একটি উপাদান x_1 এর বিন্যাসকে লেখা হয়েছে

$$f_1(x_1) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \cdots, x_p)\, dx_2\, dx_3 \cdots dx_p$$

আবার, $f_1(x_1)$-কে লেখা যায়

$$f_1(x_1) = \frac{f(x_1, x_2, \ldots, x_p)}{f_2(x_2, x_3, \cdots, x_p/x_1)}$$

অথবা $\quad f(x_1, x_2, \ldots, x_p) = f_1(x_1)\, f_2(x_2, x_3, \ldots, x_p/x_1)$

কিন্তু

$$f(x_1, x_2, \ldots, x_p) = \frac{|R|^{1/2}}{(2\pi)^{p/2}} e^{-\frac{1}{2}(x_1 - \mu_1)(R_{11} - R_{12} R_{22}^{-1} R_{21})(x_1 - \mu_1)'}$$

$$\times e^{-\frac{1}{2}\{X_2 - [M_2 - R_{22}^{-1} R_{21}(x_1 - \mu_1)]\} R_{22} \{X_2 - [M_2 - R_{22}^{-1} R_{21} \times (x_1 - \mu_1)]\}'}$$

$$= \frac{[|R_{22}||R_{11}-R_{12}R_{22}^{-1}R_{21}|^{1/2}}{(2\pi)^{1/2}(2\pi)^{\frac{1}{2}(p-1)}}$$

$$\times e^{-\frac{1}{2}(x_1-\mu_1)(R_{11}-R_{12}R_{22}^{-1}R_{21})(x_1-\mu_1)}$$

$$\times e^{-\frac{1}{2}\{X_2-[M_2-R_{22}^{-1}R_{21}(x_1-\mu_1)]\}'R_{22}\{X_2-[M_2R_{22}^{-1}R_{21}} \times (x_1-\mu_1)]\}$$

আগেই পাওয়া গিয়েছে

$$f_1(x_1) = \frac{|R_{11}-R_{12}R_{22}^{-1}R_{21}|^{1/2}}{(2\pi)^{1/2}} \times e^{-\frac{1}{2}(x_1-\mu_1)(R_{11}-R_{12}R_{22}^{-1}R_{21})(x_1-\mu_1)}$$

সুতরাং $f_2(x_2, x_3\cdots, x_p/x_1) = \frac{|R_{22}|^{1/2}}{(2\pi)^{\frac{1}{2}(p-1)}}$

$$\times e^{-\frac{1}{2}\{X_2-[M_2-R_{22}^{-1}R_{21}(x_1-\mu_1)]\}'} \times R_{22}\{X_2-[M_2-R_{22}^{-1}R_{21}(x_1-\mu_1)]\}$$

এই $f_2(\ )$ হলো শর্তাধীন বিন্যাস।

শর্তাধীন বিন্যাস একটি উপ-ভেক্টরের জন্যও নির্ণয় করা যায়, ধরা যাক

$$X=\begin{pmatrix}X_1\\X_2\end{pmatrix}, \quad \mu=\begin{pmatrix}M_1\\M_2\end{pmatrix} \text{ এবং } \Sigma=\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$$

এখানে X_1 হলো $(S\times 1)$ অর্ডারের একটি উপ-ভেক্টর $(S < p)$। Σ_{11} হলো $(S\times S)$ অর্ডারের ম্যাট্রিক্স। ধরা যাক

$$X^*=\begin{pmatrix}X_1\\X_2^*\end{pmatrix}$$

তাহলে $X_2=X_2^*$ এর ভিত্তিতে X_1 এর শর্তাধীন বিন্যাস হলো

$$h(X_1/X_2^*) = \frac{f(X_1, X_2^*)}{f_2(X_2^*)}$$

$$= \frac{(\frac{1}{2}\pi)^{p/2}|\Sigma|^{1/2}e^{-\frac{1}{2}(X^*-\mu)\Sigma^{-1}(X^*-\mu)}}{(\frac{1}{2}\pi)^{\frac{1}{2}(p-s)}|\Sigma_{22}|^{1/2}e^{-\frac{1}{2}(X_2^*-M_2)\Sigma_{22}^{-1}(X_2^*-M_2)}}$$

এখন X^*, μ ও Σ এর মান বসিয়ে পাওয়া যায় $h(X_1/X_2^*)$

$$= \frac{e^{-\frac{1}{2}[X_1 - M_1 - \Sigma_{12}\Sigma_{22}^{-1}(X_2^* - M_2)]' [\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}]^{-1}}}{(2\pi)^{s/2} |\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}|^{1/2}} \times [X_1 - M_1 - \Sigma_{12}\Sigma_{22}^{-1}(X_2^* - M_2)]$$

পরিঘাত উৎপাদক ফাংশন (Moment Generating Function) : ধরা যাক ভেক্টর X এর উপাদানসমূহের যুগ্ম পরিঘাত উৎপাদক ফাংশন হলো। $M_{X-\mu}(T)$, যেখানে

$$M_{X-\mu}(T) = E\left[\exp\left\{\sum_{j=1}^{p} t_j(x_j - \mu_j)\right\}\right] = E\left[e^{(X-\mu)'T}\right]$$

এখানে T হলো t_j উপাদানবিশিষ্ট p-মাত্রার ভেক্টর।

$$M_{X-\mu}(T)$$

$$= \frac{|R|^{1/2}}{(2\pi)^{p/2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{(X-\mu)'T - \frac{1}{2}(X-\mu)'R(X-\mu)}\, dX$$

যেখানে $dX = dx_1\, dx_2 \cdots dx_p$ ।

আবার, $(X-\mu)'T - \frac{1}{2}(X-\mu)'R(X-\mu)$

$$= -\tfrac{1}{2}(X-\mu-R^{-1}T)'R(X-\mu-R^{-1}T) + \tfrac{1}{2}T'R^{-1}T$$

$$= -\tfrac{1}{2}(X-H)'R(X-H) + \tfrac{1}{2}T'R^{-1}T,$$

যেখানে $H = \mu + R^{-1}T$

সুতরাং $M_{X-\mu}(T)$

$$= e^{\frac{1}{2}T'R^{-1}T} \frac{|R|^{1/2}}{(2\pi)^{p/2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}(X-H)'R(X-H)}\, dX$$

$$= e^{1/2T'R^{-1}T}$$

এখন R^{-1} ম্যাট্রিক্স এর (i, j) th উপাদানকে σ_{ij} ধরা হলে

$$M_{X-\mu}(T) = \exp\left(\tfrac{1}{2}\sum_i \sum_j t_i\, t_j\, \sigma_{ij}\right)$$

নিয়ামক ফাংশন (Characteristic Function) : বহুচলক পরিমিত বিন্যাসের নিয়ামক ফাংশন হলো

$$\psi(T) = E\left[\exp i \sum_{j=1}^{p} t_j x_j\right]$$

$$= \frac{|R|^{1/2}}{(2\pi)^{p/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{iX'T - \frac{1}{2}(X-\mu)'R(X-\mu)} dX$$

কিন্তু $iX'T - \frac{1}{2}(X-\mu)'R(X-\mu)$

$$= -\frac{1}{2}(X-\mu-iR^{-1}T)'R(X-\mu-iR^{-1}T) - \frac{1}{2}T'R^{-1}T$$

$$= -\frac{1}{2}(X-W)R(X-W) - \frac{1}{2}T'R^{-1}T + i\mu'T,$$

এখানে $W = \mu + iR^{-1}T$

সুতরাং,

$$\psi(T) = e^{i\mu'T - \frac{1}{2}T'R^{-1}T} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{|R|^{1/2}}{(2\pi)^{p/2}} \times e^{-\frac{1}{2}(X-W)'R(X-W)} dX$$

$$= e^{i\mu'T - \frac{1}{2}T'R^{-1}T}$$

$$= \exp\left[i \sum_{j=1}^{p} \mu_j t_j - \frac{1}{2} \sum_{i,j=1}^{p} \sigma_{ij} t_i t_j \right]$$

বহুচলক পরিমিত বিন্যাসের ধর্ম (Properties of Multivariate Normal Distribution) : ধরা যাক X হলো P-মাত্রার একটি ভেক্টর। এই X-কে P-চলক পরিমিত বিন্যাসবিশিষ্ট ভেক্টর বলা হবে, যদি এর ঘনত্ব ফাংশন (২.২.১) হয়। সেক্ষেত্রে লেখা যায় $X \sim N(\mu, \Sigma)$

উপপাদ্য ২.১ : ধরা যাক $X \sim N_p(\mu, \Sigma)$ এবং $Y = \Sigma^{-1/2}(X-\mu)$। তাহলে $y_1, y_2, \cdots y_p$ হবে অনপেক্ষ $N(0, 1)$। এখানে $\Sigma^{-1/2}$ হলো Σ^{-1} এর ধনাত্মক ডেফিনিট বর্গমূল।

প্রমাণ : $Y = \Sigma^{-1/2}(X-\mu)$ হলে

$$(X-\mu)'\Sigma^{-1}(X-\mu) = Y'Y$$

পরিবর্তনের Jacobian হবে $|\Sigma|^{1/2}$। কারণ রৈখিক পরিবর্তন $Y = AX + b$ (যেখানে A হলো একটি নন্-সিঙ্গুলার ম্যাট্রিক্স, b হলো একটি ধ্রুবকের ভেক্টর); বিবেচনা করা হলে $X = A^{-1}(Y - b)$ লেখা যায় এবং সেক্ষেত্রে $\partial x_j/\partial y_j = a^{ij}$। সুতরাং, পরিবর্তনের Jacobian হলো $\text{mod}\,|H|^{-1}$। সুতরাং, (২.২.১) অনুসারে লেখা যায়

$$g(Y) = \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}\Sigma y_i^2}$$

এখানে $\int_{-\infty}^{\infty} g(Y) = 1$। এই ঘনত্ব ফাংশন হতে বলা যায় যে, $y_1, y_2 \cdots, y_p$ হলো অনপেক্ষ $N(0, 1)$।

উপপাদ্য ২.২ ঃ ধরা যাক $X \sim N_p(\mu, \Sigma)$, তাহলে $\sum_{j=1}^{p} a_j X_j$ এর বিন্যাস হবে একচলক (univariate) পরিমিত বিন্যাস, এখানে a_j $(j = 1, 2, \cdots, p)$ হলো ধ্রুবক। এই একচলকের গড় এবং ভেদাঙ্ক হলো, যথাক্রমে $\Sigma a_j \mu_j$ এবং $\sum\sum_{j \neq k}^{p} a_j a_k \sigma_{jk}$, এখানে σ_{jk} হলো Σ এর (j, k)-তম মান।

প্রমাণ ঃ ধরা যাক $z = \Sigma a_j X_j = X'A$, যেখানে $A = (a_j)$। আমরা জানি

$$M_z(t) = E[e^{zt}] = E[e^{X'At}] = E[e^{(X-\mu)'At + \mu'At}]$$
$$= e^{\mu'At} E[e^{(X-\mu)'At}]$$

এখন বহুচলক পরিমিত বিন্যাসের পরিবাত উৎপাদক ফাংশন হতে লেখা যায়

$$M_z(t) = e^{\mu'At} e^{-\frac{1}{2}t(A'\Sigma A)t} = e^{\mu'At + \frac{1}{2}t^2(A'\Sigma A)}$$

এটি একচলক পরিমিত বিন্যাসের পরিবাত উৎপাদক ফাংশন। এখানে z এর গড় হলো $\mu'A = \Sigma a_j \mu_j$ এবং ভেদাঙ্ক হলো

$$A'\Sigma A = \sum_{j \neq k}^{p} \sum^{p} a_j a_k \sigma_{jk}$$

উপপাদ্য ২.৩ ঃ ধরা যাক $X \sim N_p(\mu, \Sigma)$। তাহলে X এর উপাদান x_j $(j = 1, 2, \cdots, p)$ গুলো যুগ্মভাবে অনপেক্ষ হবে যদি এবং কেবল যদি x_j ও x_k $(j \neq k)$ এর সহভেদাঙ্ক শূন্য হয়।

প্রমাণ ঃ x_j ও x_k এর সহভেদাঙ্ক শূন্য হলে $(j \neq k = 1, 2, \cdots, p)$ Σ ম্যাট্রিক্স হবে কৌণিক (diagonal)। সেক্ষেত্রে

$$f(x_1, x_2, \cdots, x_p) = \frac{\prod_{j=1}^{p} \sigma_{jj}^{-1/2}}{(2\pi)^{p/2}} e^{-\frac{1}{2}\sum_{j=1}^{p}(x_j-\mu_j)^2 \sigma_{jj}^{-1}}$$

এখানে $\Sigma = \text{diag}(\sigma_{jj})$। তাহলে

$$f(x_1, x_2, \cdots, x_p) = \frac{\sigma_{11}^{-1/2}}{\sqrt{2\pi}} e^{-\frac{1}{2}\frac{1}{\sigma_{11}}(x_1-\mu_1)^2}$$

$$\times \frac{\sigma_{22}^{-1/2}}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma_{22}}(x_2-\mu_2)^2} \cdots \frac{\sigma_{pp}^{-1/2}}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma_{pp}}(x_p-\mu_p)^2}$$

$$= f_1(x_1)\, f_2(x_2) \cdots f(x_p)$$

কারণ, $f_j(x_j) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_{jj}}} e^{-\frac{1}{2\sigma_{jj}}(x_j-\mu_j)^2}$

সুতরাং, Σ কৌণিক ম্যাট্রিক্স হলে x_j গুলো যুগ্মভাবে অনপেক্ষ।

এখানে দেখাতে হবে যে x_j যুগ্মভাবে অনপেক্ষ হলে Σ কৌণিক হয়। জানা আছে যে,

$$\text{Cov}(x_j, x_k) = E[(x_j-\mu_j)(x_k-\mu_k)] = \sigma_{jk},\ j \neq k$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (x_j-\mu_j)(x_k-\mu_k)\, f(x_1, x_2, \cdots, x_k)\, dx_1 \cdots dx_p$$

$$= \int_{-\infty}^{\infty} f_1(x_1)\, dx_1 \int_{-\infty}^{\infty} f_2(x_2) dx_2 \cdots \int_{-\infty}^{\infty} (x_j-\mu_j)\, f_j(x_j) dx_j$$

$$\cdots \cdots \int_{-\infty}^{\infty} (x_k-\mu_k) f_k(x_k)\, dx_k$$

[$\because$ x_j ও x_k অনপেক্ষ]

কিন্তু $$\int_{-\infty}^{\infty} (x_j-\mu_j)\, f_j(x_j) dx_j = 0$$

$\therefore$ $$\text{Cov}(x_j, x_k) = \sigma_{jk} = 0,\ j \neq k$$

উপরিউক্ত উপপাদ্য হতে বলা যায় যে, যদি X ভেক্টরকে q উপ-ভেক্টরে বিভক্ত করা যায় $(q \leqslant p)$, অর্থাৎ যদি $X = [X_1, X_2, \cdots, X_q]'$ হয় এবং $X \sim N_p(\mu, \Sigma)$ হয় তাহলে উপ-ভেক্টরগুলো যুগ্মভাবে অনপেক্ষ হবে যদি

$$f(X) = f_1(X_1)\, f_2(X_2) \cdots f_q(X_q)$$

হয়। এখানে $f_j(X_j)$ $(j = 1, 2, \cdots, q)$ হলো X_j এর j-তম প্রান্তিক বিন্যাস।

উপপাদ্য ২.৪ ঃ ধরা যাক $X \sim N_p(\mu, \Sigma)$ এবং X-কে নিম্নরূপভাবে বিভক্ত করা যায় ঃ

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix}$$

এখানে $X_1, X_2, \cdots, X_q$ হলো $(q \leqslant p)$ উপ-ভেক্টর। এই উপ-ভেক্টরগুলো যুগ্মভাবে নিরপেক্ষ হওয়ার প্রয়োজনীয় ও যথেষ্ট শর্ত হলো উপ-ম্যাট্রিক্স $\Sigma_{ij}(i \neq j)$ শূন্য ম্যাট্রিক্স এর সমান হবে।

প্রমাণ ঃ ধরা যাক $E(X_i) = M_i \qquad (i = 1, 2, \cdots, p)$। তাহলে

$$\Sigma_{ij} = E(X_i - M_i)\,(X_j - M_j)'$$

এখন $\Sigma_{ij} = 0$ $(i \neq j)$ হলে $R_{ij} = 0$ হবে, এখানে $R = \Sigma^{-1}$

সুতরাং, $f(X) = \dfrac{|R|^{1/2}}{(2\pi)^{p/2}}\, e^{-\frac{1}{2}(X-\mu)'R(X-\mu)}$

$$= \frac{|R|^{1/2}}{(2\pi)^{p/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{q}(X_i - M_i)'\,\Sigma_{ii}^{-1}(X_i - M_i)\right]$$

$$= \prod_{i=1}^{q} K_i\, e^{-\frac{1}{2}(X_i - M_i)\,\Sigma_{ii}^{-1}(X_i - M_i)}$$

$$= \prod_{i=1}^{q} f_i(X_i) = f_1(X_1)\; f_2(X_2) \cdots f_q(X_q)$$

এখানে K_i হলো ধ্রুবক।

উপরিউক্ত অনপেক্ষতা $\Sigma_{ij} = 0$ এর ক্ষেত্রে সত্য। সুতরাং দেখাতে হবে যে $\Sigma_{ij} = 0$ হবে যদি $X_1, X_2, \cdots, X_q$ অনপেক্ষ হয়।

ধরা যাক x_{ij} হলো i-তম ভেক্টরের j-তম উপাদান। সেক্ষেত্রে Σ_{ij} ম্যাট্রিক্স এর (i, k)-তম উপাদান হলো

$$E[(x_{ji}-\mu_{ji})(x_{jk}-\mu_{jk})$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(x_{ji}-\mu_{ji})(x_{jk}-\mu_{jk})f_1(X_1)f_2(X_2)\cdots$$

$$\cdots f_q(x_q)\,dX_1\cdots dX_q$$

$$=\int_{-\infty}^{\infty}f_1(X_1)\,dX_1\int_{-\infty}^{\infty}f_2(X_2)\,dX_2\ldots\int_{-\infty}^{\infty}(x_{ji}-\mu_{ji})f_i(X_i)\,dX_i\cdots$$

$$\cdots\int_{-\infty}^{\infty}(x_{jk}-\mu_{jk})f_j(X_j)\,dX_j\ \ldots\ \int_{-\infty}^{\infty}f_q(X_q)\,dX_q$$

কিন্তু i এর সকল মানের জন্য

$$\int_{-\infty}^{\infty}(x_{ji}-\mu_{ji})f_i(X_i)\,dX_i=0$$

সুতরাং, $\Sigma_{ij}=0$। এখানে $E(x_{ji})=\mu_{ji}$

উপপাদ্য ২.৫ ঃ ধরা যাক $X\sim N_p(\mu,\Sigma)$ এবং $Q=(X-\mu)'\Sigma^{-1}(X-\mu)$ হলো একটি দ্বিপদী আকার (quadratic form)। সেক্ষেত্রে μ হলো $\partial Q/\partial X=0$ এর সমাধান।

প্রমাণ ঃ $X\sim N_p(\mu,\Sigma)$ বলে

$$f(X)=\frac{|R|^{1/2}}{(2\pi)^{p/2}}e^{-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)},\quad R=\Sigma^{-1}$$

এই $f(X)$ X এর কোনো মানের জন্য সর্বোচ্চ হবে এবং $f(X)$ সর্বোচ্চ হলে $Q=(X-\mu)'\Sigma^{-1}(X-\mu)=0$ হবে। কিন্তু Σ^{-1} ধনাত্মক ডেফিনিট (positive definite) হওয়ার কারণে Q এর মান 0 হতে হলে $X-\mu=0$ হতে হবে। অর্থাৎ $X=\mu$। সুতরাং μ হলো X এর মান যা $f(X)$-কে সর্বোচ্চ করে। আবার, $\partial f(X)/\partial_X=0$ এর সমাধান থেকে প্রাপ্ত মানই $f(X)$-কে সর্বোচ্চ করে।

এই $f(X)$ অবিচ্ছিন্ন (continuous) হওয়াতে বিয়োজন করা যায়। এখন

$$\frac{f(X)}{\partial X}=\frac{|R|^{1/2}}{(2\pi)^{p/2}}e^{-\frac{1}{2}Q}\left(-\frac{1}{2}\right)\frac{\partial Q}{\partial X}=0$$

এখান থেকে বুঝা যাচ্ছে যে, ভেক্টর $\partial f(X)/\partial X = 0$ এর অন্যরূপ হলো $\partial Q/\partial X = 0$ । সুতরাং μ হলো $\partial Q/\partial X = 0$ এর সমাধান।

উপপাদ্য ২.৬ ঃ যদি $X \sim N_p(\mu, \Sigma)$ হয় এবং $Y = AX + C$ হয়, যেখানে A হলো $(q \times p)$ ম্যাট্রিক্স এবং C হলো q-মাত্রার যে কোনো ভেক্টর, তাহলে Y এর বিন্যাস হবে q-চলক পরিমিত বিন্যাস।

প্রমাণ ঃ ধরা যাক b হলো যে কোনো ধ্রুবকের q-ভেক্টর। তাহলে $b'Y = a'X + b'C$, যেখানে $a = A'b$ । কিন্তু X বহুচলক পরিমিত বিন্যাস অনুসরণ করে বলে $a'X$ হলো একচলক পরিমিত (উপপাদ্য ২.২)। সুতরাং, $b'Y$ এক-চলক পরিমিত বিন্যাস অনুসরণ করে। এটি b এর যে কোনো মানের জন্য সত্য। কাজেই Y এর বিন্যাস হবে বহুচলক পরিমিত বিন্যাস। এখন

$$E(Y) = AE(X) + C = A\mu + C$$
$$V(Y) = V(AX) = A\Sigma A'$$

উপপাদ্য ২.৭ ঃ যদি $X \sim N_p(\mu, \Sigma)$ হয়, তাহলে $(X-\mu)'\Sigma^{-1}(X-\mu)$ এর বিন্যাস হবে χ^2_p ।

প্রমাণ ঃ ধরা যাক $Y = E^{-1/2}(X-\mu)$ । এখন উপপাদ্য ২.১ অনুসারে

$$g(Y) = \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}\Sigma y_i^2}$$

অর্থাৎ ভেক্টর Y এর উপাদানসমূহ $y_1, y_2, \ldots, y_p$ যুগ্মভাবে পরিমিত বিন্যাস অনুসরণ করে। কিন্তু $g(Y)$ হতে এটিও বলা যায় যে $y_1, y_2, \ldots, y_p$ গুলো অনপেক্ষভাবে $N(0, 1)$ । সুতরাং $Y'Y = \Sigma y_i^2 \sim \chi^2_p$ । অর্থাৎ

$$Y'Y = (X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi^2_p$$

পরিমিত বিন্যাসের আরো কিছু ধর্ম (Some More Properties of Normal Distribution)

অনুসিদ্ধান্ত ২.১ ঃ যদি $X \sim N_p(0, 1)$ হয় এবং a একটি শূন্য নয় এমন p-ভেক্টর হয়, তাহলে $a'X/\sqrt{a'a}$ এর বিন্যাস হবে আদর্শ একচলক পরিমিত বিন্যাস।

অনুসিদ্ধান্ত ২.২ ঃ যদি $X \sim N_p(\mu, \sigma^2 I)$ হয় এবং $B\ (q \times p)$ অর্ডারের একটি সারি-অর্থোনরমাল ম্যাট্রিক্স হয়, যেখানে $BB' = I_q$, তাহলে $BX \sim N_q(B\mu, \sigma^2 I)$ ।

অনুসিদ্ধান্ত ২.৩ ঃ যদি $X \sim N_p(\mu, \sigma^2 I)$ হয় এবং $B(q \times p)$ অর্ডারের অর্থোনরমাল ম্যাট্রিক্স হয়, তাহলে BX এবং $(I - B'B)X$ অনপেক্ষ হবে।

অনুসিদ্ধান্ত ২.৪ ঃ যদি $X \sim N_p(\mu, \Sigma)$ হয়, তাহলে AX এবং BX অনপেক্ষ হবে যদি এবং কেবল যদি $A\Sigma B' = 0$ হয়।

পরিমিত উপাত্ত ম্যাট্রিক্সসমূহের পরিবর্তন (Transformation of Normal Data Matrics) ঃ ধরা যাক $x_1, x_2, \ldots, x_n$ হলো $N_p(\mu, \Sigma)$ হতে প্রাপ্ত n আকারের একটি নমুনা। মনে করা যাক $X = (x_1, x_2, \ldots, x_n)'$ হলে $N_p(\mu, \Sigma)$ হতে প্রাপ্ত একটি উপাত্ত ম্যাট্রিক্স। এই উপাত্ত ম্যাট্রিক্স-এর একটি রৈখিক ফাংশন বিবেচনা করা যাক $Y = A \times B$; এখানে A ও B হলো, যথাক্রমে $(m \times n)$ এবং $(p \times q)$ অর্ডারের বাস্তব মানভিত্তিক দুটি ধ্রুবক ম্যাট্রিক্স। এখানে উপাত্ত ম্যাট্রিক্স X-কে পরিবর্তন করে Y পাওয়া গিয়েছে। এরূপ একটি গুরুত্বপূর্ণ পরিবর্তন হলো $\bar{X}' = n^{-1} 1'X$, যেখানে $A = n^{-1} 1'$, $B = I_p$।

X যদি $(n \times p)$ আকারের ম্যাট্রিক্স হয় এবং $X_1', X_2', \ldots, X_n'$ X ম্যাট্রিক্স এর n সারি হয় এবং সেগুলো অনপেক্ষভাবে $N_p(\mu, \Sigma)$ হয়, তাহলে

$$f(X) = (2\pi\Sigma)^{-n/2} \exp\{-\tfrac{1}{2}\mathrm{tr}[\Sigma^{-1}(X - 1\mu')'(X - 1\mu')\}$$

এখানে X এর বিন্যাসকে ম্যাট্রিক্স পরিমিত বিন্যাস বলা হয়।

অনুসিদ্ধান্ত ২.৫ ঃ যদি $X(n \times p)$ আকারের উপাত্ত ম্যাট্রিক্স হয় এবং তা $N_p(\mu, \Sigma)$ এর একটি নমুনা হয়, তাহলে $n\bar{X} = X'1$ এর ক্ষেত্রে $\bar{X} \sim N_p(\mu, n^{-1}\Sigma)$ হবে।

অনুসিদ্ধান্ত ২.৬ ঃ যদি X $N_p(\mu, \Sigma)$ হতে $(n \times p)$ আকারের একটি উপাত্ত ম্যাট্রিক্স হয় এবং $Y = A \times B$ পরিবর্তন বিবেচনা করা হয়, তাহলে Y একটি পরিমিত উপাত্ত ম্যাট্রিক্স হবে, যদি এবং কেবল যদি

(i) $A1 = \alpha 1$ অথবা $B'\mu = 0$ হয়, এখানে α হলো একটি ধ্রুবক।

(ii) $A'A = \beta I$ অথবা $B'\Sigma B = 0$, এখানে β হলো একটি ধ্রুবক।

উপরিউক্ত শর্ত দুটি পূরণ হলে $Y \sim N_q(\alpha B'\mu, \beta B'\Sigma B)$।

অনুসিদ্ধান্ত ২.৭ ঃ যদি X $N_p(\mu, \Sigma)$ হতে একটি উপাত্ত ম্যাট্রিক্স হয় এবং যদি $Y = A \times B$ ও $Z = C \times D$ পরিবর্তন বিবেচনা করা হয়, তাহলে Y ও Z এর উপাদানসমূহ অনপেক্ষ হবে, যদি এবং কেবল যদি

(i) $B'\Sigma D = 0$। অথবা (ii) $AC' = 0$ হয়।

অনুসিদ্ধান্ত ২.৮ ঃ যদি X $N_p(\mu, \Sigma)$ হতে একটি উপাত্ত ম্যাট্রিক্স হয় এবং $\bar{X} = n^{-1}X'1$; HX অনপেক্ষ হয়, তাহলে $\bar{X}$ ও $S = n^{-1}X'HX$ অনপেক্ষ হবে।

অনুসিদ্ধান্ত ২.৯ ঃ যদি X $N_p(\mu, \Sigma)$ হতে একটি উপাত্ত ম্যাট্রিক্স হয় এবং X-কে যদি $X=(X_1\ X_2)$ দুটি উপ-ভেক্টরে বিভক্ত করা যায়, তাহলে X_1 ও $X_{2\cdot1}=X_2-X_1\Sigma_{11}^{-1}\Sigma_{12}$ হবে অনপেক্ষ। তাছাড়া X_1 ও $X_{2\cdot1}$ যথাক্রমে $N(\mu_1,\Sigma_{11})$ এবং $N(\mu_{2\cdot1},\Sigma_{22\cdot1})$ হতে উপাত্ত ম্যাট্রিক্স হবে। এখানে $E(X_1)=\mu_1$, $E(X_{2\cdot1})=\mu_{2\cdot1}=\mu_2-\Sigma_{21}\Sigma_{11}^{-1}\mu_1$, $\Sigma_{22\cdot1}=\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$

প্রমাণ ঃ ধরা যাক $X=(X_1\ X)$, $\Sigma=\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$, $\mu=\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}$

মনে করি $B'=[I\ 0]$ এবং $D'=(-\Sigma_{21}\Sigma_{11}^{-1}, I)$ তাহলে $X_1=XB$ এবং $X_{2\cdot1}=XD$। তাছাড়া $B'\Sigma D=0$। সুতরাং অনুসিদ্ধান্ত ২.৭ অনুসারে বলা যায় যে, X_1 ও $X_{2\cdot1}$ অনপেক্ষ।

সিঙ্গুলার বহুচলক পরিমিত বিন্যাস (Singular Multivariate Normal Distribution) ঃ X $N_p(\mu, \Sigma)$ হলে $f(X)$-এ Σ^{-1} বিদ্যমান। কিন্তু $r(\Sigma)=k<p$ হলে Σ^{-1} পাওয়া যায় না। সেক্ষেত্রে Σ হলো সিঙ্গুলার ম্যাট্রিক্স। অবশ্য Σ এর জেনারালাইজড্ উল্টো ম্যাট্রিক্স (g-inverse) পাওয়া যায়। ধরা যাক $\lambda_1, \lambda_2, \cdots, \lambda_k$ হলো Σ এর নিয়ামক মূল। তাহলে, X-এর বিন্যাস হবে সিঙ্গুলার বহুচলক পরিমিত বিন্যাস এবং এর ঘনত্ব ফাংশন (p. d. f) হলো

$$\frac{(2\pi)^{-k/2}}{(\lambda_1\lambda_2\cdots\lambda_k)^{1/2}}\exp\left\{-\frac{1}{2}(X-\mu)'\Sigma^{-}(X-\mu)\right\}$$

এখানে Σ^{-} হলো g-inverse, X ভেক্টর হাইপারপ্লেন (Hyperplane) $N'(X-\mu)=0$-এ বিদ্যমান। N হলো $[P(P-K)$ ম্যাট্রিক্স, $N'\Sigma=0$ এবং $N'N=I_{p-k}$

২.৩ Wishart বিন্যাস (Wishart Distribution)

ধরা যাক X হলো $(n\times p)$ আকারের ম্যাট্রিক্স। তাহলে $X'CX$ হলো দ্বিপদী আকার (Quadratic form) এবং এটি একটি প্রতিসম ম্যাট্রিক্স। এই $X'CX$ এর বিন্যাস হলো Wishart বিন্যাস। ধরা যাক X ম্যাট্রিক্স হলো $N_p(\mu, \Sigma)$ হতে চয়ন করা নমুনা ম্যাট্রিক্স। সেক্ষেত্রে $X'CX$ হবে নমুনা বর্গসমষ্টি ও গুণন-সমষ্টি এর ম্যাট্রিক্স যদি $C=\left[I-\frac{1}{n}11'\right]$ হয়। এই ম্যাট্রিক্স $X'CX$-কে বলা হয় Wishart ম্যাট্রিক্স।

ধরা যাক $A=(X-1\mu')'(X-1\mu')=(a_{ij})$

একটি ম্যাট্রিক্স $(i, j = 1, 2, \cdots, p)$। এই ম্যাট্রিক্স-এ $p(p+1)/2$ ভিন্ন মান আছে। এখানে A-এর বিন্যাস হবে এর ভিন্ন মানগুলোর যুগ্ম বিন্যাস এবং এটিকে বলা হয় Wishart বিন্যাস এবং এর ঘনত্ব ফাংশন হলো

$$W_p = K(n, p)\,|A|^{\frac{n-p-1}{2}}\, e^{-\frac{1}{2}\mathrm{tr}\,A}, \quad A > 0$$

এখানে $1/K(n, p) = 2^{np/2}\,\pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma\left(\frac{n+1-i}{2}\right)$

উপাত্ত ম্যাট্রিক্স X এর বিন্যাসের ভিত্তিতে A ম্যাট্রিক্স-এর রূপ বিভিন্ন হয়। যেমন,

(i) $A = X'X$, যদি $X \sim N_p(0, I)$ বা $X \sim N_p(0, \Sigma)$

(ii) $A = (X - 1\mu')'(X - 1\mu')$, যদি $X \sim N_p(\mu, \Sigma)$

এখানে μ এর পরিবর্তে নমুনা গড় ভেক্টর $\bar{X}$ ব্যবহার করেও A এর সংজ্ঞায়ন করা যায়। সেক্ষেত্রে

$$A = (X - 1\bar{X}')'(X - 1\bar{X}'), \text{ যেখানে}$$

$$\bar{X} = n^{-1}1'X = n^{-1}X'1 = n^{-1}\sum_{i=1}^{n} X_i$$

Wishart বিন্যাসকে $W_p(\Sigma, n)$ দ্বারা চিহ্নিত করা হয়ে থাকে। অর্থাৎ $A \sim W_p(\Sigma, n)$, এখানে Σ-কে বলা হয় মাপনী পরামান (scale parameter) এবং n-কে বলা হয় স্বাধীনতার মাত্রা (degrees of freedom) পরামান। যদি $X \sim N_p(0, I)$ হয়, তাহলে A এর বিন্যাসকে আদর্শ (standard) Wishart বিন্যাস বলা হয়।

Wishart বিন্যাসের ব্যুৎপত্তি (Derivation of Wishart Distribution) : ধরা যাক $X_1, X_2, \cdots, X_p$ হলো p ভেক্টর। এই ভেক্টরসমূহকে Gram-Schmidt পদ্ধতির মাধ্যমে সমকৌণিকতার পরিবর্তন করা যাক, যেখানে

$$Y_i = b^{i1}X_1 + b^{i2}X_2 + \cdots + b^{ii}X_i \; ; \; i = 1, 2, \cdots, p$$

এই $b^{i1}, b^{i2}, \cdots, b^{ii}$ এর মান এমনভাবে নিতে হবে যেন $Y_1, Y_1, Y_2, \cdots Y_{i-1}$ এর সমকৌণিক হয় এবং Y_i এর দৈর্ঘ্য (length) এক হয়। এখন সমকৌণিক পরিবর্তনকে লেখা যাক

$Y = XB^{-1}$, এখানে

$$B^{-1} = \begin{pmatrix} b^{11} & 0 & 0 & \cdots & 0 \\ b^{21} & b^{22} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \\ b^{p1} & b^{p2} & \cdots & & b^{pp} \end{pmatrix}$$

এবং

$$B = \begin{pmatrix} b_{11} & 0 & 0 & \cdots & 0 \\ b_{21} & b_{22} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \\ b_{p1} & b_{p2} & \cdots & & b_{pp} \end{pmatrix}$$

$\therefore \quad X = YB$

ধরা যাক Wishart ম্যাট্রিক্স হলো $A = X'X$। তাহলে

$$A = B'Y'YB = B'B \quad [\because Y'Y = I_p]$$

এই ম্যাট্রিক্স B-কে বলা হয় A এর Bartlett বিযোজন (decomposition)। এবং লেখা যায়

$$X_j = b_{1j}Y_1 + b_{2j}Y_2 + \cdots + b_{jj}Y_j$$

একে Y_j' দ্বারা প্রাক্-গুণন করে পাওয়া যায়

$$Y_j'X_j = b_{1j}Y_j'Y_1 + b_{2j}Y_j'Y_2 + \cdots + b_{jj}Y_j'Y_j$$
$$= b_{jj} \quad [\because Y_j'Y_1 = 0 \; ; \; Y_j'Y_j = 1]$$
$$j = 1, 2, \cdots, i \; ; \; i = 1, 2, \cdots, p$$

আবার $X_i'X_i = b_{i1}^2 Y_1'Y_1 + b_{i2}^2 Y_2'Y_2 + \ldots + b_{ii}^2 Y_i'Y_i$

$$\therefore \quad b_{ii}^2 = X_i'X_i - \sum_{j=1}^{i-1} b_{ij}^2$$

$$= a_{ii} - \sum_{j=1}^{i-1} b_{ij}^2 \quad [\because A = X'X]$$

এখানে সমকৌণিক পরিবর্তনের মাধ্যমে X_1 থেকে $b_{i1}, b_{i2}, \cdots, b_{ii-1}$ পাওয়া গিয়েছে। আবার $x_{ij}(i=1, 2, \cdots, p\ ;\ j=1, 2, \cdots, n)$ $N(0, I)$ হতে চয়ন করা নমুনা বিবেচনা করা হলে $b_{i1}, b_{i2}, \cdots, b_{ii-1}$ এর বিন্যাসও হবে $N(0, I)$। সুতরাং,

$$b_{ii}^2 = \left(X_i' X_i - \sum_{j=1}^{i-1} b_{ij}^2\right) \sim \chi^2_{n-(i-1)}$$

কাজেই B এর বিন্যাস হবে

$$f(B)\, dB = \prod_{i=1}^{p} \prod_{j=1}^{i-1} \left\{ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} b_{ij}^2} d\, b_{ij} \right\}$$

$$\times \prod_{k=1}^{p} \left\{ \chi^2_{n-(k-1)},\ b_{kk}^2\, d(b_{kk}^2) \right\}$$

$$-\infty < b_{ij} < \infty\ ; \quad 0 < b_{ii} < \infty$$

এখন B এর বিন্যাস হতে A এর বিন্যাস নির্ণয় করতে হবে। জানা আছে

$$A = B'B \Rightarrow |A| = \prod_{i=1}^{p} b_{ii}^2$$

পরিবর্তনের Jacobian হলো

$$J(B \to A) = \frac{1}{J(A \to B)} = 2^p \prod_{i=1}^{p} (b_{ii})^{1-p-1}$$

আবার $$\operatorname{tr} A = \operatorname{tr} B'B = \sum_{i=1}^{p} \sum_{j=1}^{i} b_{ij}^2$$

সুতরাং, A এর p.d.f. হলো

$$W_p(I, n) = K(n, p)\, |A|^{\frac{1}{2}(n-p-1)} e^{-\frac{1}{2}\operatorname{tr} A}, \quad A > 0$$

এখানে $K(n, p) = 2^{\frac{1}{2}np}\pi^{\frac{1}{4}p(p-1)} \prod_{i=1}^{p} \left\{\Gamma\left(\frac{n+1-i}{2}\right)\right\}$

সাধারণ Wishart বিন্যাস (General Wishart Distribution) : ধরা যাক $X(n \times p)$ হলো উপাত্ত ম্যাট্রিক্স যা $N_p(\mu, \Sigma)$ হতে চয়ন করা নমুনা। উক্ত উপাত্ত ম্যাট্রিক্স-এর ভিত্তিতে Wishart ম্যাট্রিক্স ধরা যাক

$$A = (X - 1\mu')'(X - 1\mu')$$

এখন একটি পরিবর্তন $Y = (X-\mu)C^{-1}$ বিবেচনা করা যাক, যেখানে $\Sigma = C'C$ এবং C হলো ট্রায়াঙ্গুলার ম্যাট্রিক্স। তাহলে $Y \sim N_p(0, 1)$ এবং

$$D = Y'Y = (C^{-1})'(X-\mu)'(X-\mu)C^{-1} = (C^{-1})'AC^{-1}$$

এর বিন্যাস হবে Wishart বিন্যাস। এখানে D থেকে A-তে পরিবর্তন করতে হবে। পরিবর্তনের Jacobian হবে

$$J(D \to A) = |C^{-1}|^{p+1} = |\Sigma|^{-\frac{1}{2}(p+1)}$$

আবার $|D| = |(C^{-1})'|\,|A|\,|C^{-1}| = |A|/|\Sigma|$

এবং $\mathrm{tr}D = \mathrm{tr}(C^{-1})'AC^{-1} = \mathrm{tr}(C'C)^{-1}A = \mathrm{tr}\Sigma^{-1}A$

$$\therefore \quad f(A) = \frac{K(n, p)}{|\Sigma|^{n/2}} |A|^{\frac{1}{2}(n-p-1)} e^{-\frac{1}{2}\mathrm{tr}\Sigma^{-1}A}, \quad A > 0$$

μ এর পরিবর্তে $\bar{X}$ ব্যবহার করে Wishart বিন্যাস (Wishart Distribution using $\bar{X}$ in place of μ) : ধরা যাক উপাত্ত ম্যাট্রিক্স $X(n \times p)$ হলো $N_p(\mu, \Sigma)$ হতে চয়ন করা n আকারের নমুনা। নমুনা উপাত্তসমূহের সম্ভাব্যতা (likelihood) ফাংশন হলো

$$L = \frac{1}{(2\pi)^{\frac{1}{2}np}|\Sigma|^{n/2}} e^{-\frac{1}{2}\mathrm{tr}\,\Sigma^{-1}(X-1\mu')'(X-1\mu')}$$

$$-\infty < X < \infty$$

এক্ষেত্রে $A = (X - 1\bar{X}')'(X - 1\bar{X}')$ হলো Wishart ম্যাট্রিক্স। ধরা যাক μ এবং Σ এর সর্বোচ্চ সম্ভাব্য নিরূপক (ML estimator) হলো যথাক্রমে

$$\hat{\mu} = \bar{X} = n^{-1}1'X = n^{-1}X'1$$

এবং $$\hat{\Sigma} = S + (\bar{X} - \mu)'(\bar{X} - \mu)$$

এখানে $S = (X - 1\bar{X}')'(X - 1\bar{X}')$

এখন X ম্যাট্রিক্স হতে Y ম্যাট্রিক্স-এ একটি পরিবর্তন করা যাক, যেখানে $Y = PX$, এখানে P হলো সমকৌণিক ম্যাট্রিক্স এবং এর শেষ সারি হলো $n^{-1/2}1'$। পরিবর্তনের Jacobian হলো

$$J(X \to Y) = |P|^p = 1, \quad [\because P \text{ হলো সমকৌণিক}]$$

যেহেতু P এর শেষ সারি হলো $n^{-1/2}1'$, Y-কে নিম্নরূপভাবে বিভক্ত করা যায় :

$$Y = \begin{bmatrix} Z \\ \sqrt{n}\,\bar{X} \end{bmatrix}$$

এখানে Z হলো $(n-1) \times P$ অর্ডারের ম্যাট্রিক্স।

আবার, $PI = [0, 0, \cdots, 0, \sqrt{n}]'$

কারণ P এর প্রথম $(n-1)$ সারি শেষ সারির সমকৌণিক। সুতরাং

$$(X - 1\mu')'(X - 1\mu') = (X - 1\mu')'P'P(X - 1\mu')$$

$$= (Y - P1\mu')'(Y - P1\mu')$$

$$= \begin{bmatrix} Z \\ \sqrt{n}(\bar{X} - \mu) \end{bmatrix}' \begin{bmatrix} Z \\ \sqrt{n}(\bar{X} - \mu) \end{bmatrix}$$

$$= Z'Z + n(\bar{X} - \mu)'(\bar{X} - \mu)$$

$$-\infty < Z < \infty; \quad -\infty < \sqrt{n}\,\bar{X} < \infty$$

তাহলে $L = \dfrac{1}{(2\pi)^{\frac{1}{2}np}\,|\Sigma|^{n/2}}\, e^{-\frac{1}{2}\mathrm{tr}\,\Sigma^{-1}[Z'Z + n(\bar{X} - \mu)'(\bar{X} - \mu)]}$

উপরিউক্ত ফাংশনকে দুটি উপাদানে বিভক্ত করা যায়, যার একটি উপাদান হলো $Z'Z$ বিশিষ্ট এবং অপরটি হলো $n(\bar{X} - \mu)'(\bar{X} - \mu)$ বিশিষ্ট। সুতরাং বলা যায় যে Z এবং $\sqrt{n}\,\bar{X}$ অনপেক্ষভাবে বিন্যাসিত। তাছাড়া

$$\sqrt{n}\,\bar{X} \sim N_p(\sqrt{n}\,\mu, \Sigma)$$

আবার সম্ভাব্যতা ফাংশন হতে বলা যায় যে, Z হলো P-মাত্রার বহুচলক পরিমিত বিন্যাস হতে চয়ন করা $(n-1)$ আকারের নমুনা এবং $E(Z) = 0$ ও $V(Z) = \Sigma$। সুতরাং $Z'Z \sim W_p(\Sigma, n-1)$। কিন্তু P সমকৌণিক হওয়াতে

$$Z'Z = Y'Y - n\bar{X}'\bar{X} = X'P'PX - n\bar{X}'X$$

$$= X'[I - n^{-1}11']X = S$$

সুতরাং, $S \sim W_p(\Sigma, n-1)$ ।

Wishart বিন্যাসের নিয়ামক ফাংশন (Characteristic Function of Wishart Distribution) : ধরা যাক উপাত্ত ম্যাট্রিক্স $X(n \times p)$ $N_p(\mu, \Sigma)$ হতে চয়ন করা n আকারের নমুনা। মনে করি $A = (a_{ij}) = X'X$ হলো Wishart-ম্যাট্রিক্স। এই ম্যাট্রিক্সে $a_{11}, a_{12}, \cdots, a_{pp}$ নামক $P(P+1)/2$ মান আছে। এখন A এর নিয়ামক ফাংশন নির্ণয় করার মানে হলো উক্ত $P(P+1)/2$ ভিন্ন ভিন্ন মানের জন্য নিয়ামক ফাংশন নির্ণয় করতে হবে। ধরা যাক $\psi(T)$ হলো প্রয়োজনীয় নিয়ামক ফাংশন। তাহলে,

$$\psi(T) = \psi(t_{11}, t_{12}, \cdots, t_{1p}, t_{21}, t_{22}, \cdots, t_{2p}, \cdots, t_{pp})$$

$$= E[\exp\{i(t_{11}a_{11} + t_{12}a_{12} + \cdots + a_{pp}t_{pp})\}]$$

$$= E\left[\exp\left\{\frac{i}{2}\mathrm{tr}AK\right\}\right]$$

এখানে $t_{ij} = t_{ji}$, $i > j$, $K_{ij} = (1 + \delta_{ij})t_{ij}$, $K = (K_{ij})$

δ_{ij} হলো Kronecker ডেল্টা।

$$\therefore \quad \psi(T) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{\frac{1}{2}i\mathrm{tr}AK}\frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \times e^{-\frac{1}{2}\mathrm{tr}\Sigma^{-1}X'X}dX$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \times e^{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1} - iK)X'X}dX \quad [\because A = X'X]$$

$$= \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}(2\pi)^{np/2}|\Sigma^{-1} - iK|^{-n/2}$$

$$= \frac{|\Sigma^{-1}|^{-n/2}|I - iK\Sigma|^{-n/2}}{|\Sigma^{-1}|^{-n/2}} = \frac{1}{|I - iK\Sigma|^{n/2}}$$

Wishart বিন্যাসের ধর্ম (Properties of Wishart Distribution) :

উপপাদ্য ২.৮ ঃ ধরা যাক $A = X'X \sim W_p(n, \Sigma)$ এবং B হলো $(p \times q)$ ম্যাট্রিক্স। তাহলে $B'AB \sim W_q(B'\Sigma B, n)$ ।

প্রমাণ ঃ ধরা যাক X হলো $(n \times p)$ অর্ডারের উপাত্ত ম্যাট্রিক্স যা $N_p(0, \Sigma)$ হতে চয়ন করা নমুনা। সে কারণে $A = X'X \sim W_p(n, \Sigma)$ । একটি পরিবর্তন $Y = XB$ বিবেচনা করা যাক। তাহলে $Y \sim N_q(0, B'\Sigma B)$। এখন, $B'AB = B'X'XB = Y'Y$ । কিন্তু $Y'Y \sim W_q(n, B'\Sigma B)$

সুতরাং, $B'AB \sim W_q(n, B'\Sigma B)$

অনুসিদ্ধান্ত ২.১০ : যদি $A \sim W_p(n, I)$ এবং B একটি $(p \times q)$ অর্ডারের ম্যাট্রিক্স হয় এবং $B'B = I_q$ হয়, তাহলে $B'AB \sim W_q(n, I)$ ।

উপপাদ্য ২.৯ ঃ যদি $A \sim W_p(n, \Sigma)$ হয় এবং b একটি p-মাত্রার ভেক্টর হয়, তাহলে $u = b'Ab/b'\Sigma b$ এর বিন্যাস হবে χ^2-বিন্যাস।

প্রমাণ ঃ উপপাদ্য ২.৮ হতে বলা যায় $b'Ab \sim W_1(n, b'\Sigma b)$ । সুতরাং $u = b'Ab/b'\Sigma b$ এর বিন্যাস χ^2_n ।

অনুসিদ্ধান্ত ২.১১ ঃ যদি $A \sim W_p(n, \Sigma)$ হয়, তাহলে $|A| / |\Sigma|$ এর বিন্যাস হবে p অনপেক্ষ কাই-বর্গ বিন্যাসের গুণ যে গুলোর স্বাধীনতার মাত্রা হবে যথাক্রমে $n, n-1, \cdots n-(p-1)$ ।

অনুসিদ্ধান্ত ২.১২ ঃ যদি $A \sim W_p(n, \Sigma)$ হয়, তাহলে σ^{pp}/a^{pp} এর বিন্যাস হবে $n-(p-1)$ স্বাধীনতার মাত্রাবিশিষ্ট কাই-বর্গ বিন্যাস। এখানে σ^{pp} এবং a^{pp} হলো, যথাক্রমে Σ^{-1} ও A^{-1} এর সর্বশেষ মান।

অনুসিদ্ধান্ত ২.১৩ ঃ যদি $X(n \times p)$ অর্ডারের উপাত্ত ম্যাট্রিক্স হয় এবং এটি $N_p(0, \Sigma)$ হতে চয়ন করা নমুনা হয় তাহলে $X'CX$ এর বিন্যাস হবে Wishart বিন্যাস যদি এবং কেবল যদি $C(n \times n)$ অর্ডারের প্রতিসম আইডেমপোটেন্ট ম্যাট্রিক্স হয়। সেক্ষেত্রে $X'CX \sim W_p(r, \Sigma)$, এখানে r হলো C ম্যাট্রিক্স-এর পদসংখ্যা (rank)।

অনুসিদ্ধান্ত ২.১৪ ঃ যদি $A_1, A_2, \cdots, A_k$ $(p \times p)$ অর্ডারের k ম্যাট্রিক্স হয়, যেখানে $A_i \sim W_p(n_i, \Sigma)$ $(i = 1, 2, \cdots, k)$, তাহলে ম্যাট্রিক্স

$$A = \sum_{i=1}^{k} A_i \sim W_p(n, \Sigma) \; ; \text{ যেখানে } n = \sum_{i=1}^{k} n_i$$

অনুসিদ্ধান্ত ২.১৫ ঃ যদি X ম্যাট্রিক্স-এর সারিসমূহ অনপেক্ষ হয় এবং একই $N_p(\mu, \Sigma)$ অনুসরণ করে এবং যদি $C_1, C_2, \cdots, C_k$ প্রতিসম ম্যাট্রিক্স হয়,

তাহলে $X'C_1X, X'C_2X, \cdots, X'C_kX$ যুগ্মভাবে অনপেক্ষ হবে যদি $C_iC_j = 0$ $(i \neq j)$ হয়।

অনুসিদ্ধান্ত ২.১৬ ঃ যদি $A \sim W_p(n, \Sigma)$ হয়, তাহলে $E(A) = n\Sigma$ এবং $E(A^{-1}) = (n-p-1)^{-1}\Sigma^{-1}$, যদি $(n-p-1) > 0$ হয়।

উপপাদ্য ২.১০ ঃ যদি $A \sim W_p(n, \Sigma)$ এবং $B \sim W_p(m, \Sigma)$ অনপেক্ষ হয় এবং যদি $n \geq p$, $m \geq p$ হয়, তাহলে

$$\varphi = |A^{-1}B| = |B| / |A|$$

p অনপেক্ষ F চলকের গুণের সমানুপাতিক হবে, যেখানে i-তম চলকের স্বাধীনতার মাত্রা হলো $(m-i+1)$ এবং $(n-i+1)$।

প্রমাণ ঃ অনুসিদ্ধান্ত ২.১১ হতে বলা যায় যে, $|A|$ এবং $|B|$ হলো p অনপেক্ষ $|\Sigma| \chi^2$ বিন্যাস। সুতরাং φ হলো p অনপেক্ষ χ^2 তথ্যভমানের অনুপাতের গুণ। কাজেই i-তম অনুপাত হলো

$$(m-i+1)/(n-i+1)F_{m-i+1,\, n-i+1}$$

এটি i-এর সকল মানের জন্য সত্য।

অনুসিদ্ধান্ত ২.১৭ ঃ যদি $A \sim W_p(m, I)$ এবং $B \sim W_p(n, I)$ অনপেক্ষ হয় $(m \geq p)$, তাহলে $\Lambda = |A| / |A+B| = |I + A^{-1}B|$ এর বিন্যাস হবে Wilk's ল্যামডা $[\Lambda(p, m, n)]$ বিন্যাস। আবার,

$$\Lambda(p, m, n) = \prod_{i=1}^{n} u_i, \text{ যেখানে } u_1, u_2, \cdots, u_n$$

হলো অনপেক্ষ এবং $u_i \sim B(\frac{1}{2}(m+i-p, \frac{1}{2}p))$ চলক $(i = 1, 2, \cdots, n)$

বিভক্ত Wishart ম্যাট্রিক্সসমূহ (Partitioned Wishart Matrices) : ধরা যাক $A \sim W_p(m, \Sigma)$। এই A-কে অনেক সময় কতকগুলো উপ-ম্যাট্রিক্স সমূহে বিভক্ত করার প্রয়োজন হয়। যেমন, উপাত্ত ম্যাট্রিক্স X ($n \times p$) অর্ডারের হলে p-চলকসমূহকে দুটি গুচ্ছে বিভক্ত করা যেতে পারে। ধরা যাক প্রথম গুচ্ছে a সংখ্যক এবং দ্বিতীয় গুচ্ছে b সংখ্যক চলক আছে, এখানে $a+b=p$। এখন চলক গুচ্ছে বিভক্তকরণ প্রসঙ্গে A ম্যাট্রিক্সকেও নিম্নরূপভাবে বিভক্ত করা যেতে পারে :

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

এখানে A_{11} ও A_{22} এর অর্ডার হলো যথাক্রমে $(a \times a)$ এবং $(b \times b)$ । এখানে A_{11}, A_{12}, A_{21} এবং A_{22} হলো বিভক্ত Wishart ম্যাট্রিক্সসমূহ । A_{11} ও A_{22} এর বিন্যাস হলো Wishart বিন্যাস । এই বিভক্ত Wishart ম্যাট্রিক্সসমূহ হতে লেখা যায়

$$A_{22\cdot 1} = A_{22} - A_{21} A_{11}^{-1} A_{12}$$

এখানে A_{11} ও $A_{22\cdot 1}$ অনপেক্ষ ।

এখন $A \sim W_p(m, \Sigma)$, $m > a$ হলে

(i) $A_{22\cdot 1} \sim W_b(m-a, \Sigma_{22\cdot 1})$ এবং এটি A_{11} ও A_{12} এর অনপেক্ষ ।

(ii) যদি $\Sigma_{12} = 0$ হয়, তাহলে $A_{22} - A_{22\cdot 1} = A_{21} A_{11}^{-1} A_{12}$ এর বিন্যাস হবে $W_b(a, \Sigma_{22})$ । তাছাড়া $A_{21}A_{11}^{-1}A_{12}$, A_{11} এবং $A_{22.1}$ যুগ্মভাবে অনপেক্ষ ।

২.৪ Hotelling T^2 বিন্যাস (Hotelling T^2 Distribution)

ধরা যাক X হলো $(n \times p)$ অর্ডারের উপাত্ত ম্যাট্রিক্স যা $N_p(0, I)$ হতে চয়ন করা নমুনা । তাহলে $A = X'X$ এর বিন্যাস হলো $W_p(n, I)$ । এখানে X ও A এর বিন্যাস অনপেক্ষ । তাহলে $nX'A^{-1}X$ বিন্যাস হলো Hotelling T^2 বিন্যাস, যার দুটি পরামান হলো p এবং n । একে লেখা হয় $nX'A^{-1}X \sim T^2(p, n)$ ।

আবার, ধরা যাক Y এবং M অনপেক্ষভাবে, যথাক্রমে $N_p(\mu, \Sigma)$ ও $W_p(m, \Sigma)$ বিন্যাস অনুসরণ করে, তাহলে $m(Y-\mu)'M^{-1}(Y-\mu) \sim T^2(p, m)$ । কারণ, উপপাদ্য ২.১ অনুসারে $M^{-1/2}(Y-\mu) \sim N_p(0, I)$ । আবার, $\Sigma^{-1/2} M \Sigma^{-1/2} \sim W_p(m, I)$ । সুতরাং প্রথমোক্ত সংজ্ঞা অনুসারে $m(Y-\mu)'M^{-1}(Y-M) \sim T^2(p, m)$ ।

উপরিউক্ত ফলাফলের ভিত্তিতে এটিও বলা যায় যে,

$$(n-1)(\bar{Y}-\mu)'S_S^{-1}(\bar{Y}-\mu) = n(\bar{Y}-\mu)'S^{-1}(\bar{Y}-\mu) \sim T^2(p, n-1)$$

এখানে $\bar{Y}$ এবং S হলো যথাক্রমে $N_p(\mu, \Sigma)$ হতে n আকারের নমুনা ভিত্তিক গড় ভেক্টর ও সহ-ভেদাঙ্ক ম্যাট্রিক্স ।

আগেই লক্ষ্য করা গিয়েছে যে, $S_S = \frac{n}{n-1} S$ । কাজেই উপরে আলোচিত Hotelling T^2 এর দ্বিতীয় ফলাফলের ক্ষেত্রে $M = nS$ বসিয়ে, $M^{-1/2}(Y-\mu)$ এর পরিবর্তে $n^{1/2}M^{-1/2}(\bar{Y}-\mu)$ এবং $m = n-1$ বসিয়ে পাওয়া যায় যে,

$$(n-1)(\bar{Y}-\mu)'S_S^{-1}(\bar{Y}-\mu) \sim T^2(p, n-1) \text{ ।}$$

আলোচিত T^2-তথ্যজমানের (statistic) উদ্ভব করেন Hotelling (1931)। এর উৎপত্তি হলো নাস্তিকল্পনা $H_0 : \mu = \mu_0$ যাচাই করার জন্য। অবশ্য অন্যান্য আরো নাস্তিকল্পনা যাচাই করার জন্য T^2-তথ্যজমানের ব্যবহার করা হয়। কিন্তু $H_0 : \mu = \mu_0$ যাচাই করার জন্য সম্ভাব্যতা অনুপাত যাচাই (likelihood ratio test) ব্যবহার করা হলেও এর উদ্ভব হয়। বিষয়টি নিরীক্ষা করে দেখা যাক।

ধরা যাক $X_1, X_2, \cdots, X_n$ $(n>p)$ হলো $N_p(\mu, \Sigma)$ হতে চয়ন করা n আকারের নমুনা। এই নমুনা তথ্যমানগুলোর সম্ভাব্যতা ফাংশন (likelihood function) হলো

$$L(\mu, \Sigma^{-1}) = \frac{|\Sigma^{-1}|^{n/2}}{(2\pi)^{np/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)'\Sigma^{-1}(X_i-\mu)\right]$$

উপরিউক্ত L এর ভিত্তিতে সম্ভাব্যতা অনুপাত নির্দেশক (criterion) হলো

$$\lambda = \frac{\max L(\mu_0, \Sigma^{-1})}{\max L(\mu, \Sigma^{-1})}$$

এখানে λ এর লব (Numerator) হলো $\mu = \mu_0$ এর ভিত্তিতে L এর সর্বোচ্চ মান এবং হর (Denominator) হলো μ ও Σ^{-1} এর ভিত্তিতে L এর সর্বোচ্চ মান, যেখানে Σ^{-1} হলো ধনাত্মক ডেফিনিট ম্যাট্রিক্স। μ ও Σ^{-1} এর জন্য কোনো শর্ত আরোপ করা না হলে, L এর সর্বোচ্চ মান হবে μ ও Σ^{-1} এর সর্বোচ্চ সম্ভাব্যতা নিরূপকের ক্ষেত্রে। এখন μ ও Σ এর সর্বোচ্চ সম্ভাব্যতা নিরূপক (Maximum likelihood estimate) হলো, যথাক্রমে

$$\hat{\mu} = \bar{X} \quad \text{এবং} \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})'$$

কিন্তু $\mu = \mu_0$ এর ক্ষেত্রে

$$\hat{\Sigma}_H = \frac{1}{n}\Sigma(X_i-\mu_0)(X_i-\mu_0)'$$

এখন L-এ $\hat{\mu}$ ও $\hat{\Sigma}$ এর মান বসিয়ে এবং সরল করে পাওয়া যায়

$$\max L(\mu, \Sigma^{-1}) = \frac{1}{(2\pi)^{np/2}|\hat{\Sigma}|^{n/2}} e^{-\frac{1}{2}np}$$

$$\max L(\mu_0, \Sigma^{-1}) = \frac{1}{(2\pi)^{np/2}|\hat{\Sigma}_H|^{n/2}} e^{-\frac{1}{2}np}$$

তাহলে, $$\lambda = \frac{|\hat{\Sigma}|^{n/2}}{|\hat{\Sigma}_H|^{n/2}} = \frac{|\Sigma(X_i-\bar{X})(X_i-\bar{X})'|^{n/2}}{|\Sigma(X_i-\mu_0)(X_i-\mu_0)'|^{n/2}}$$

$$= \frac{|A|^{n/2}}{|A+n(\bar{X}-\mu_0)(\bar{X}-\mu_0)'|^{n/2}}$$

এখানে $$A = \sum_i (X_i-\bar{X})(X_i-\bar{X})' = (n-1)S_0$$

কাজেই $$\lambda^{2/n} = \frac{|A|}{|A+[\sqrt{n}(\bar{X}-\mu_0)][\sqrt{n}(\bar{X}-\mu_0)]'|}$$

$$= \frac{|A|}{\begin{vmatrix} 1 & \sqrt{n}(\bar{X}-\mu_0)' \\ -\sqrt{n}(\bar{X}-\mu_0) & A \end{vmatrix}}$$

$$= \frac{1}{1+T^2/(n-1)} \qquad (২.৪.১)$$

$$\left[\because \begin{vmatrix} B & C \\ D & E \end{vmatrix} = \begin{vmatrix} B & C \\ D & E \end{vmatrix} \cdot \begin{vmatrix} I & -B^{-1}C \\ 0 & I \end{vmatrix} \right.$$

$$= \begin{vmatrix} B & 0 \\ D & E-DB^{-1}C \end{vmatrix} = |B| \cdot |E-DB^{-1}C|$$

যদি $$|B| \neq 0 \,]$$

এখানে

$$T^2 = n(\bar{X}-\mu_0)' S^{-1}(\bar{X}-\mu_0) = (n-1)n(\bar{X}-\mu_0)'A^{-1}(\bar{X}-\mu_0)$$

T^2 এর বিন্যাস (Distribution of T^2) : ধরা যাক $T^2 = Y'S^{-1}Y$, যেখানে $Y \sim N_p(\gamma, \Sigma)$ এবং $nS \sim W_p(n, \Sigma)$ অনপেক্ষভাবে বিন্যাসিত, এখানে $nS = \Sigma X_i X_i'$ । X_i গুলো অনপেক্ষ এবং প্রতিটি $N_p(0, \Sigma)$ বিন্যাস অনুসরণ করে । ধরা যাক D হলো একটি নন্-সিঙ্গুলার ম্যাট্রিক্স এবং $D\Sigma D' = I$ । এখন একটি পরিবর্তন বিবেচনা করা যাক

$$Z = DY$$

তাহলে,
$$S_z = DSD'$$
$$\gamma_z = D\gamma$$

উক্ত পরিবর্তিত চলকের ভিত্তিতে $T^2 = Z' S_z^{-1} Z$, যেখানে $Z \sim N(\gamma_z, I)$ এবং nS_z অনপেক্ষভাবে বিন্যাসিত। এখানে $nS_z = \Sigma DX_i (DX_i)'$ এবং DX_i এর প্রতিটি বিন্যাস হলো $N(0, I)$। লক্ষ্য করা যাচ্ছে যে,

$$\gamma' \Sigma^{-1} \gamma = \gamma_z' (I)^{-1} \gamma_z = \gamma_z' \gamma_z$$

এখন একটি $(p \times p)$ সমকৌণিক ম্যাট্রিক্স C বিবেচনা করা যাক, যেখানে C এর প্রথম সারির মানগুলো হবে

$$C_{1i} = \frac{Z_i}{\sqrt{Z'Z}}, \quad i = 1, 2, \cdots, p$$

যার ফলে $\Sigma C_{1i}^2 = 1$। ম্যাট্রিক্স C এর অন্যান্য $(p-1)$ সারির মানগুলো হবে এমন যেন $CC' = I$। যেহেতু C ম্যাট্রিক্স Z এর উপর নির্ভরশীল, সে কারণে C হলো দৈব ম্যাট্রিক্স। এখন ধরা যাক

$$U = CZ, \quad B = CnS_zC'$$

এখান থেকে পাওয়া যায়

$$U_1 = \Sigma C_{1i} Z_i = \sqrt{Z'Z}$$

$$U_j = \Sigma C_{ji} Z_i$$
$$= \sqrt{Z'Z}\ \Sigma C_{ji} C_{1i} = 0 \qquad , \qquad j \neq 1$$

তাহলে

$$\frac{T^2}{n} = U' B^{-1} U = (U_1\ 0\ 0 \cdots 0) \begin{pmatrix} b^{11} & b^{12} & \cdots & b^{1p} \\ b^{21} & b^{22} & \cdots & b^{2p} \\ \cdots & \cdots & & \cdots \\ b^{p1} & b^{p2} & \cdots & b^{pp} \end{pmatrix} \begin{pmatrix} U_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$= U_1^2 b^{11}$$

এখানে $(b^{ij}) = B^{-1}$। জানা আছে যে,

$$1/b^{11} = b_{11} - b_{(1)} B_{22}^{-1} b'_{(1)} = b_{11,2,\cdots,p}$$

এখানে $b_{(1)}=[b_{12}\ b_{13}\dots b_{1p}]$, $B_{22}=\begin{bmatrix} b^{22} & \dots & b^{2p} \\ \vdots & & \vdots \\ b^{p2} & \dots & b^{pp} \end{bmatrix}$

এবং $B=\begin{bmatrix} b_{11} & b_{(1)} \\ b'_{(1)} & b_{22} \end{bmatrix}$

সুতরাং, $T^2/n = U_1^2/b_{11\cdot 2,\dots,p} = Z'Z/b_{11\cdot 2,\dots,p}$

ধরা যাক $V_1 = CX_1$ হলো অনপেক্ষ এবং V_1 এর প্রতিটির বিন্যাস হলো $\Sigma V_1 V_1'$। কাজেই C দেয়া থাকলে B এর শর্তাধীন বিন্যাস হবে $\Sigma V_1 V_1'$ এর বিন্যাস। সুতরাং $b_{11\cdot 2,\dots,p}$ শর্তাধীনে $n-(p-1)$ স্বাধীনতার মাত্রাবিশিষ্ট χ^2 বিন্যাস। কিন্তু $b_{11\cdot 2,\dots,p}$ এর শর্তাধীন বিন্যাস C এর উপর নির্ভর করে না বিধায় এর নিঃশর্ত বিন্যাসও χ^2। আবার, $Z'Z$ এর বিন্যাস হলো p স্বাধীনতার মাত্রাবিশিষ্ট অকেন্দ্রিক (noncentral) χ^2 বিন্যাস যার অকেন্দ্রিকতার পরামান (noncentrality parameter) হলো $\gamma_Z'\gamma_Z = \gamma'\Sigma^{-1}\gamma$। সুতরাং T^2/n এর বিন্যাস হলো একটি অকেন্দ্রিক χ^2 ও একটি অনপেক্ষ χ^2 বিন্যাসের অনুপাত।

উপরিউক্ত আলোচনা হতে বলা যায় যে,

$$T^2 = \frac{np}{n-p+1} F_{p,\ n-p+1} \qquad (২.৪.২)$$

বা $\left[\frac{T^2}{n}\right]\frac{n-p+1}{p}$ এর বিন্যাস হলো p এবং $(n-p+1)$ স্বাধীনতার মাত্রাবিশিষ্ট অকেন্দ্রিক F বিন্যাস যার অকেন্দ্রিকতার পরামান হলো $\gamma'\Sigma^{-1}\gamma$।

অনুসিদ্ধান্ত ২.১৮ : যদি $\bar{Y}$ ও S যথাক্রমে $N_p(\mu, \Sigma)$ হতে চয়ন করা n আকারের নমুনার গড় ভেক্টর ও সহ-ভেরাঙ্ক ম্যাট্রিক্স হয়, তাহলে

$$\{(n\dots p)/p\}\{(\bar{Y}-\mu)'S^{-1}(\bar{Y}-\mu) \sim F_{p,\ n-p}$$

দুই নমুনার ভিত্তিতে Hotelling T^2 তথ্যজমান (Hotelling T^2 Statistic Based on Two Samples): ধরা যাক X_1 হলো $N_p(\mu_1, \Sigma)$ হতে n_1 আকারের এবং X_2 হলো $N_p(\mu_2, \Sigma)$ হতে n_2 আকারের দুটি উপাত্ত ম্যাট্রিক্স। এক্ষেত্রে পরামান ভেক্টরের জন্য Mahalanobis দূরত্ব এর সংজ্ঞায়ন করা হয়েছে (২.১.১)

$$\Delta^2 = (\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$$

ধরা যাক $n_1 + n_2 = n$ এর উভয় নমুনার ভিত্তিতে নমুনা সহ-ভেদাঙ্ক ম্যাট্রিক্স হলো $S_s = (n_1S_1 + n_2S_2)/(n-2)$ এবং এটি Σ এর নিখুঁতি নিরূপক। আরো ধরা যাক যে $\bar{X}_i$ এবং $S_i\,(i = 1, 2)$ হলো যথাক্রমে নমুনা গড় ভেক্টর এবং নমুনা সহ-ভেদাঙ্ক ম্যাট্রিক্স। তাহলে দুই নমুনার জন্য নমুনা Mahalanobis দূরত্ব হবে

$$D^2 = (\bar{X}_1 - \bar{X}_2)' S_s^{-1}(\bar{X}_1 - \bar{X}_2)$$

এই D^2 এর তথ্যজমান হলো Hotelling $T^2(p, n-2)$। এটি একটি উপপাদ্যের মাধ্যমে প্রমাণ করা যাক।

উপপাদ্য ২.১১ ঃ যদি X_1 ও X_2 দুটি অনপেক্ষ উপাত্ত ম্যাট্রিক্স হয় এবং যদি X_i এর n_i সারিসমূহ অনপেক্ষভাবে ও একইরূপে $N_p(\mu_i, \Sigma_i)$ বিন্যাস অনুসরণ করে, তাহলে $\mu_1 = \mu_2$ এবং $\Sigma_1 = \Sigma_2$ এর ক্ষেত্রে $(n_1n_2/n)\,D^2 = T^2(p, n-2)$

প্রমাণ : $X_i\,(i = 1, 2) \sim N_p(\mu_i, \Sigma_i)$ হওয়াতে $\bar{X}_i \sim N_p(\mu_i, n_i^{-1}\Sigma_i)$। কাজেই $\bar{X}_1 - \bar{X}_2$ এর বিন্যাসও পরিমিত বিন্যাস, যার গড় ভেক্টর এবং সহ-ভেদাঙ্ক ম্যাট্রিক্স হলো, যথাক্রমে $\mu_1 - \mu_2$ এবং $n_1^{-1}\Sigma_1 + n_2^{-1}\Sigma_2$। আবার $\mu_1 = \mu_2$ এবং $\Sigma_1 = \Sigma_2$ হলে $\bar{X}_1 - \bar{X}_2 \sim N_p(0, C\Sigma)$; এখানে $C = n/n_1n_2\ (n = n_1 + n_2)$।

ধরা যাক $A_i = nS_i$, তাহলে $A_i \sim W_p(n_i - 1, \Sigma_i)$। এখন $\Sigma_1 = \Sigma_2$ হলে $A_1 + A_2 = A \sim W_p(n-2, \Sigma)$। এখানে $A = (n-2)S_s$। আবার $(n/n_1n_2)A \sim W_p(n-2, C\Sigma)$ এবং $A\,(\bar{X}_1 - \bar{X}_2)$ এর অনপেক্ষ। কারণ $\bar{X}_i$ ও $S_i\,(i = 1,2)$ অনপেক্ষভাবে বিন্যাসিত। সুতরাং

$$(n-2)\,(\bar{X}_1 - \bar{X}_2)' \left(\frac{n}{n_1n_2} A\right)^{-1} (\bar{X}_1 - \bar{X}_2) \sim T^2\,(p, n-2)$$

জানা আছে $D^2 = (\bar{X}_1 - \bar{X}_2)' S_s^{-1}(\bar{X}_1 - \bar{X}_2)$

সুতরাং $\dfrac{n_1n_2}{n} D^2 = \dfrac{n_1n_2}{n}\,(\bar{X}_1 - \bar{X}_2)' S_s^{-1}(\bar{X}_1 - \bar{X}_2)$

কিন্তু $S_s = (n_1S_1 + n_2S_2)/(n-2)$ Σ এর নিখুঁতি নিরূপক

$= A/(n-2)$ হওয়াতে

$$\frac{n_1n_2}{n} D^2 = T^2\,(p, n-2)$$

অনুসিদ্ধান্ত ২.১৯ ঃ (২.৪.২) এর ভিত্তিতে লেখা যায়

$$\frac{n_1 n_2 (n-p-1)}{n(n-2)\,p} D^2 \sim F_{p,\, n-p-1}$$

Mahalanobis দূরত্ব এর বিযোজন (Decomposition of Mahalanobis Distance) : ধরা যাক X হলো $(n \times p)$ অর্ডারের উপাত্ত ম্যাট্রিক্স যা $N_p(0,\Sigma)$ হতে চয়ন করা নমুনা, Y হলো p-মাত্রার একটি ভেক্টর, যেখানে $Y \sim N_p(\mu,\Sigma)$। আরো বিবেচনা করা যাক যে X ও Y অনপেক্ষভাবে বিন্যাসিত। মনে করি $A = X'X \sim W_p(m, \Sigma)$; তাহলে নমুনাভিত্তিক Mahalahobis দূরত্ব হবে

$$D_p^2 = mY'A^{-1}Y$$

একে বিভক্ত করে লেখা যায়

$$D_p^2 = D_k^2 + mZ'A_{22.1}^{-1}Z$$

এখানে

$$D_k^2 = m\, Y_1' A_{11}^{-1} Y_1,\ A_{22\cdot 1} = A_{22} - A_{21} A_{11}^{-1} A_{12},\ Z = Y_2 - A_{21} A_{11}^{-1} Y_1$$

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \quad \text{এবং} \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$$

এখানে Y_1 হলো K-মাত্রার ভেক্টর। Y এর গড় ভেক্টর μ-কে বিভক্ত করে লেখা যায় $\mu' = (\mu_1', \mu_2')$, যেখানে μ_1 হলো প্রথম k চলকের গড় উপ-ভেক্টর। এখন Σ-কে বিভক্ত করে পাওয়া যায় $\Sigma_{22\cdot 1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$। আবার, $\mu_{2\cdot 1} = \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1$। সুতরাং μ ও 0 এর মধ্যে Mahalanobis দূরত্বকে বিযোজন করে লেখা যায়

$$\begin{aligned} \Delta_p^2 &= \mu'\Sigma^{-1}\mu \\ &= \mu_1'\Sigma_{11}^{-1}\mu_1 + \mu_{2.1}'\Sigma_{22.1}^{-1}\mu_{2\cdot 1} \\ &= \Delta_k^2 + \mu_{2.1}'\Sigma_{22.1}^{-1}\mu_{2\cdot 1} \end{aligned}$$

এখানে

Δ_k^2 হলো প্রথম k চলকের ভিত্তিতে Mahalanobis দূরত্ব। $\Delta_k^2 = \Delta_p^2$ হলে, $\mu_{2\cdot 1} = 0$ হবে।

উপরিউক্ত আলোচনা এবং D^2 এর বিন্যাস হতে বলা যায় যে,

$$\frac{D_p^2 - D_k^2}{m + D_k^2} \sim \frac{p-k}{m-p+1} F_{p-k,\, m-p+1} \qquad (২.৪.৩)$$

এবং এটি D_k^2 এর অনপেক্ষ।

২.৫ Wilks এর ল্যাম্বডা বিন্যাস (Wilks' Lambda Distribution)

অনুসিদ্ধান্ত ২.১৭ এর মাধ্যমে Wilks' ল্যাম্বডা বিন্যাসের সংজ্ঞায়ন করা হয়েছে। সম্ভাব্যতা অনুপাত যাচাই (likelihood ratio test) এর ক্ষেত্রে এই বিন্যাসের ব্যবহার প্রায়ই লক্ষ্য করা যায়। নিচে একটি উপপাদ্য-এর মাধ্যমে Λ এর বিন্যাস আলোচনা করা হলো।

উপপাদ্য ২.১২ ঃ ধরা যাক $y_1, y_2, \ldots, y_n$ হলো n অনপেক্ষ বিটা চলক এবং i-তম চলকের পরামান হলো $y_i \sim B[\frac{1}{2}(m+i-p), \frac{1}{2}p]$; $i=1, 2, \ldots, n$। তাহলে y_i গুলোর গুণফলের বিন্যাস হবে $\Lambda(p, m, n)$। অর্থাৎ

$$\Lambda(p, m, n) \sim \prod_{i=1}^{n} y_i$$

প্রমাণ ঃ ধরা যাক X হলো $(n \times p)$ অর্ডারের উপাত্ত ম্যাট্রিক্স যা $N_p(0, I)$ হতে চয়ন করা নমুনা। আরো ধরা যাক যে X_i হলো $(i \times p)$ অর্ডারের ম্যাট্রিক্স যা X ম্যাট্রিক্স এর প্রথম i সারিসমূহ দ্বারা গঠিত। এখন $M = X'X$ হলে

$$A_i = B + X_i'X_i, \quad i = 1, 2, \ldots, n$$

আরো লেখা যায়

$$A_i = A_{i-1} + x_i x_i'$$

কাজেই $\quad A_0 = B, \quad A_n = B + M$

এখন
$$\Lambda(p, m, n) = \frac{|B|}{|B+M|} = \frac{|A_0|}{|A_n|}$$

$$= \frac{|A_0|}{|A_1|}\frac{|A_1|}{|A_2|}\cdots\frac{|A_{n-1}|}{|A_n|}$$

$$= y_1, y_2, \ldots, y_n$$

যেখানে
$$y_i = \frac{|A_{i-1}|}{|A_i|}, \quad i = 1, 2, \cdots, n$$

কিন্তু জানা আছে যে যদি $d \sim N_p(0, I)$ এবং $A \sim W_p(m, I)$ হয়, তাহলে $|A| / |A+dd'| \sim B[\frac{1}{2}(m-p+1), \frac{1}{2}p]$। কাজেই উক্ত ফলাফলের ভিত্তিতে বলা যায় যে $y_i \sim B[\frac{1}{2}(m+i-p), \frac{1}{2}p]$, $i = 1, 2, \cdots, n$। আলোচিত y_i গুলো অনপেক্ষও। কারণ, A_i

$$1 + x_i'A_{i-1}^{-1}x_i = |A_i| / |A_{i-1}| = y_i^{-1}$$

এর অনপেক্ষ। এখন y_i $x_{i+1}, x_{i+2}, \ldots x_n$ এর অনপেক্ষ হওয়াতে এবং

$$A_{i+1} = A_i + \sum_{k=1}^{j} x_{i+k} x'_{i+k}$$

হওয়াতে y_i $A_{i+1}, A_{i+2}, \ldots, A_n$ এর অনপেক্ষ। এখানে x_i হলো X ম্যাট্রিক্স এর i-তম সারির মান দ্বারা গঠিত p-মাত্রার ভেক্টর।

উপপাদ্য ২.১৩ ঃ $\Lambda(p, m, n)$ এবং $\Lambda(n, m+n-p, p)$ বিন্যাস দুটি একই।

প্রমাণ ঃ জানা আছে যে $A \sim W_p(m, I)$ এবং $B \sim W_p(n, I)$ অনপেক্ষ হলে $\Lambda = |A| / |A+B| = |I - A^{-1}B|^{-1} \sim \Lambda(p, m, n)$। ধরা যাক $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p$ হলো $A^{-1}B$ এর নিয়ামক মূল। মনে করি $K = \min(n, p)$ হলো শূন্য নয় এমন নিয়ামক মূলের সংখ্যা। তাহলে নিয়ামক মূলের ধর্ম হতে লেখা যায়

$$\Lambda(p, m, n) = |I + A^{-1}B|^{-1} = \prod_{i=1}^{p} (1+\lambda_i)^{-1} = \prod_{i=1}^{k} (1+\lambda_i)^{-1}$$

অর্থাৎ Λ হলো $A^{-1}B$ এর নিয়ামক মূলের ফাংশন। আবার জানা আছে $m \geqslant p$ এবং $n, p \geqslant 1$ হলে এবং $\psi(p, m, n)$ ও $\psi(n, m+n-p, p)$ এর শূন্য নয় এমন নিয়ামক মূলের সংখ্যা সমান হলে $\psi(p, m, n)$ ও $\psi(n, m+n-p, p)$ একই বিন্যাস। উক্ত ফলাফলের ভিত্তিতে বলা যায় যে $\Lambda(p, m, n)$ ও $\Lambda(n, m+n-p, p)$ এর বিন্যাস একই।

তৃতীয় অধ্যায়

নিরূপণ ও যাচাই পদ্ধতি
(Method of Estimation and Test)

৩.১ নিরূপণ পদ্ধতি (Method of Estimation)

ধরা যাক p-চলক পরিমিত গণসমষ্টি হতে n আকারের একটি দৈব নমুনা আছে। নমুনা তথ্যমানের ভিত্তিতে সম্ভাব্যতা ফাংশন (likelihood function) হলো

$$L = |2\pi\Sigma|^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)'\Sigma^{-1}(X_i-\mu)\right\}$$

এখন μ এবং Σ-কে এমনভাবে নিরূপণ করতে হবে যেন L সর্বোচ্চ হয়। এখানে

$$\log L = -\frac{n}{2}\log|2\pi\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)'\Sigma^{-1}(X_i-\mu)$$

$$= -\frac{n}{2}\log|2\pi\Sigma|$$

$$-\frac{1}{2}\left[\sum_{i=1}^{n}(X_i-\bar{X})'\Sigma^{-1}(X_i-\bar{X})+n(\bar{X}-\mu)'\Sigma^{-1}(\bar{X}-\mu)\right]$$

কিন্তু $(X_i-\bar{X})'\Sigma^{-1}(X_i-\bar{X})$ স্কেলার (scaler) হওয়াতে একে এর trace হিসেবে লেখা যায়। ফলে

$$\log L = -\frac{n}{2}\log|2\pi\Sigma|$$

$$-\frac{1}{2}\left[\mathrm{tr}\,\Sigma^{-1}\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})'+n(\bar{X}-\mu)'\Sigma^{-1}(\bar{X}-\mu)\right]$$

ধরা যাক $\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})' = nS$। তাহলে

$$\log L = -\frac{n}{2}\log|2\pi\Sigma| - \frac{n}{2}\operatorname{tr}\Sigma^{-1}S - \frac{n}{2}(\bar{X}-\mu)'\Sigma^{-1}(\bar{X}-\mu) \qquad (৩.১.১)$$

এখন μ ও Σ এর নিরূপক পাওয়া যাবে, যথাক্রমে সমীকরণ

$$\frac{\partial \log L}{\partial \mu} = 0 \quad \text{ও} \quad \frac{\partial \log L}{\partial \Sigma^{-1}} = 0$$

হতে। ধরা যাক $V = \Sigma^{-1}$, তাহলে

$$\log L = -\frac{n}{2}\log 2\pi + \frac{n}{2}\log|V| - \frac{n}{2}\operatorname{tr} VS$$

$$-\frac{n}{2}\operatorname{tr} V(\bar{X}-\mu)(\bar{X}-\mu)'$$

$[\because \quad (\bar{X}-\mu)'\Sigma^{-1}(\bar{X}-\mu)$ হলো স্কেলার $]$

এখন logL-কে বিযোজন করার জন্য নিচের ফলাফল ব্যবহার করা যেতে পারে।

$X(n \times p)$ এর প্রসঙ্গে $f(X)$ এর বিযোজন ফলকে ম্যাট্রিক্স আকারে ধরা যাক

$$\frac{\partial f(X)}{\partial X} = \left(\frac{\partial f(X)}{\partial x_{ij}}\right)$$

তাহলে,

(i) $\dfrac{\partial a'X}{\partial X} = a$, (ii) $\dfrac{\partial X'X}{\partial X} = 2X$,

(iii) $\dfrac{\partial X'AX}{\partial X} = (A + A')X$, (iv) $\dfrac{\partial X'AY}{\partial X} = AY$

(v) $\dfrac{\partial |X|}{\partial x_{ij}} = X_{ij}$, যদি $X(n \times p)$-এর সকল মান ভিন্ন ভিন্ন হয়।

$$= \begin{cases} X_{ij}, & i = j \\ 2X_{ij} & i \neq j \end{cases} \} \; X \text{ যদি প্রতিসম হয়।}$$

এখানে X_{ij} হলো X-এর (i, j)-তম সহ-পূরক (co-factor)

(vi) $\dfrac{\partial \operatorname{tr} XY}{\partial X} = Y'$, যদি $X(n \times p)$-এর সকল মান ভিন্ন ভিন্ন হয়।

$= Y + Y' - \operatorname{diag}(Y)$, যদি $X(n \times p)$ প্রতিসম হয়।

(vii) $\dfrac{\partial X^{-1}}{\partial x_{ij}} = -X^{-1}J_{ij}X^{-1}$, যদি $X(n \times p)$-এর সকল মান ভিন্ন ভিন্ন হয়।

$$= -X^{-1}J_{ij}X^{-1}, \quad i=j$$
$$= -X^{-1}(J_{ij}+J_{ji})X^{-1}, \quad i\neq j \Big\} \text{, যদি } X \text{ প্রতিসম হয়।}$$

এখানে J_{ij} হলো এমন একটি ম্যাট্রিক্স (i, j)-তম স্থানে 1 এবং অন্য স্থানে শূন্য।

উপরিউক্ত ফলাফলের ভিত্তিতে পাওয়া যায়

$$\frac{\partial \log L}{\partial \mu} = n(\bar{X}-\mu)=0 \rightarrow \bar{X}=\hat{\mu}$$

এখন $\frac{\partial \log L}{\partial V}$ এর মান নির্ণয় করতে $\log L$ এর ডান পাশের প্রতিটি রাশিকে ভিন্ন ভিন্ন ভাবে বিবেচনা করা যাক। তাহলে

$$\frac{\partial}{\partial v_{ij}}\left|\log V\right| = \begin{cases} 2V_{ij}/|V|, & i\neq j \\ V_{ii}/|V|, & i=j \end{cases}$$

এখানে V_{ij} হলো V ম্যাট্রিক্স এর (i, j)-তম সহ-পূরক। কিন্তু V প্রতিসম হওয়ার কারণে, $V_{ij}/|V|$ ম্যাট্রিক্স হলো $V^{-1}=\Sigma$। সুতরাং,

$$\frac{\partial \log |V|}{\partial V} = 2\Sigma - \text{diag } \Sigma$$

আবার, $\quad \frac{\partial \text{tr } VS}{\partial V} = 2S - \text{diag } S \quad (\because S \text{ প্রতিসম})$

এবং $\quad \frac{\partial \text{tr } V(\bar{X}-\mu)(\bar{X}-\mu)'}{\partial V} = 2(\bar{X}-\mu)(\bar{X}-\mu)' - \text{diag}(\bar{X}-\mu)(\bar{X}-\mu)'$

$$\therefore \quad \frac{\partial \log L}{\partial V} = \frac{n}{2}\big[\, 2\Sigma - \text{diag } \Sigma - 2S + \text{diag } S$$
$$-2(\bar{X}-\mu)(\bar{X}-\mu)' + \text{diag } (\bar{X}-\mu)(\bar{X}-\mu)' \,\big]$$
$$= \frac{n}{2}\big[\, 2M - \text{diag } M \,\big]$$

এখানে $\quad M = \Sigma - S - (\bar{X}-\mu)(\bar{X}-\mu)$

$\frac{\partial \log L}{\partial V}=0$ বসিয়ে পাওয়া যায়

$$2M - \text{diag } M = 0 \rightarrow M = 0$$

$$\therefore \quad \Sigma - S - (\bar{X}-\mu)(\bar{X}-\mu)' = 0$$

বা $$\hat{\Sigma} = S + (\bar{X} - \hat{\mu})(\bar{X} - \hat{\mu})'$$

কিন্তু $\hat{\mu} = \bar{X}$ হওয়াতে $\hat{\Sigma} = S$

উদাহরণ ১.২-এর ক্ষেত্রে M. Gagates এর দৈহিক ওজন ও দৈহিক দৈর্ঘ্য এর ভিত্তিতে গড় ভেক্টর এবং সহ-ভেদাঙ্ক ম্যাট্রিক্স পাওয়া যায় নিম্নরূপ :

$$\bar{X}[\,3.23 \quad 41.07\,]', \quad S = \begin{bmatrix} 1.7076 & 1.0711 \\ 1.0711 & 7.2622 \end{bmatrix}$$

এখানে μ ও Σ এর সর্বোচ্চ সম্ভাব্য নিরূপক (MLE) হিসেবে, যথাক্রমে $\bar{X}$ ও S-কে বিবেচনা করা যায়।

μ জানা থাকলে Σ নিরূপণ (Estimation of Σ when μ is known) : ধরা যাক μ এর মান জানা আছে এবং $\mu = k\,\mu_0$। এখন Σ-ও জানা থাকলে

$$\log L = -\frac{n}{2}\log |\,2\pi\Sigma\,| - \frac{n}{2}\operatorname{tr}\Sigma^{-1}S - \frac{n}{2} \times (\bar{X} - k\mu_0)\Sigma^{-1}(\bar{X} - k\mu_0)'$$

তাহলে $$\frac{\partial \log L}{\partial k} = -n\mu_0'\Sigma^{-1}(\bar{X} - k\mu_0) = 0$$

$$\rightarrow \hat{k} = \mu_0'\Sigma^{-1}\bar{X}/\mu_0'\Sigma^{-1}\mu_0$$

কিন্তু Σ অজানা হলে Σ ও k এর নিরূপক পাওয়ার জন্য সমীকরণ হলো

$$\hat{\Sigma} = S + (\bar{X} - \mu)(\bar{X} - \mu)' \qquad \text{(a)}$$

$$\hat{k} = \mu_0'\Sigma^{-}\bar{X}/\mu_0'\Sigma^{-1}\mu_0 \qquad \text{(b)}$$

এখন (a)-কে $\hat{\Sigma}^{-1}$ দ্বারা প্রাক-গুণন এবং S^{-1} দ্বারা পর-গুণন করে পাওয়া যায়

$$S^{-1} = \hat{\Sigma}^{-1} + \hat{\Sigma}^{-1}(\bar{X} - \mu)(\bar{X} - \mu)'\,S^{-1}$$

একে μ_0 দ্বারা প্রাক্-গুণন করে এবং $n\mu_0'\Sigma^{-1}(\bar{X} - k\mu_0) = 0$ বসিয়ে পাওয়া যায়

$$\mu_0'S^{-1} = \mu_0'\hat{\Sigma}^{-1} \Rightarrow = S^{-1} = \hat{\Sigma}^{-1} \rightarrow \hat{\Sigma} = S$$

$$\therefore \quad \hat{k} = \mu_0' S^{-1} \bar{X} / \mu_0' S^{-1} \mu_0$$

ধরা যাক উপরে আলোচিত উদাহরণের ক্ষেত্রে

$\mu = [\ 3.00 \quad 41.00\ ]$। তাহলে

$$\hat{\Sigma} = S + (\bar{X} - \mu)(\bar{X} - \mu)'$$

$$= \begin{bmatrix} 1.7076 & 1.0711 \\ 1.0711 & 7.2622 \end{bmatrix} + \begin{bmatrix} 0.23 \\ 0.07 \end{bmatrix} \begin{bmatrix} 0.23 & 0.07 \end{bmatrix}$$

$$= \begin{bmatrix} 1.7555 & 1.0872 \\ 1.0872 & 7.2671 \end{bmatrix}$$

কিন্তু $\mu = k\mu_0 = k[3.00 \quad 41.00]$ অনুমান করা হলে

$$\hat{k} = \mu_0' S^{-1} \bar{X} / \mu_0' S^{-1} \mu_0$$

এখানে $S^{-1} = \begin{bmatrix} 0.6453 & -0.0952 \\ -0.0952 & 0.1517 \end{bmatrix}$

$$\mu_0' S^{-1} \bar{X} = [\ 3.00 \quad 41.00\ ] \begin{bmatrix} 0.6453 & -0.0952 \\ -0.0952 & 0.1517 \end{bmatrix} \begin{bmatrix} 3.23 \\ 41.07 \end{bmatrix}$$

$$= 237.36$$

$$\mu_0' S^{-1} \mu_0 = [\ 3.00 \quad 41.00\ ] \begin{bmatrix} 0.6453 & -0.0952 \\ -0.0952 & 0.1517 \end{bmatrix} \begin{bmatrix} 3.00 \\ 41.00 \end{bmatrix}$$

$$= 237.40$$

$$\therefore \quad \hat{k} = \mu_0' S^{-1} \bar{X} / \mu_0' S^{-1} \mu_0 = 0.9998$$

সুতরাং $\hat{\mu}' = \hat{k}\mu_0' = 0.9998\ [\ 3.00 \quad 41.00\ ]'$

$$= [\ 2.9994 \quad 40\ ]$$

Σ-এর উপর শর্ত আরোপ করে নিরূপণ পদ্ধতি (Method of Estimation under Constraint on Σ) : ধরা যাক Σ এর জন্য শর্ত দেয়া আছে $\Sigma = k\Sigma_0$; তাহলে (৩.১.১) হতে লেখা যায়

$$2n^{-1}\log L = -p\log k - \log|2\pi\Sigma_0| = k^{-1}a, \text{ এখানে}$$

$a = \text{tr}\,\Sigma_0^{-1}S + (\bar{X}-\mu)'\Sigma_0^{-1}(\bar{X}-\mu)$। এখন μ জানা থাকলে $\frac{\partial \log L}{\partial k} = 0$ হতে k এর মান পাওয়া যায়। এখানে

$$\frac{\partial \log L}{\partial k} = -p/k + a/k^2 = 0 \rightarrow \hat{k} = a/p$$

কিন্তু μ এর মান দেয়া না থাকলে k ও μ এর মান নির্ণয় করতে হবে

$$\frac{\partial \log L}{\partial k} = 0 \text{ এবং } \frac{\partial \log L}{\partial u} = 0$$

সমীকরণদ্বয় হতে। উক্ত সমীকরণদ্বয় হতে পাওয়া যায় $k = a/p$, $\hat{\mu} = \bar{X}$। উক্ত দুটি মান হতে পাওয়া যায় $\hat{k} = \text{tr}\,\Sigma_0^{-1}S/p$।

এরূপ নিরূপণে $\Sigma_{12} = 0$ শর্ত আরোপ করা থাকলে Σ-এর নিরূপক হবে

$$\hat{\Sigma} = \begin{bmatrix} S_{11} & 0 \\ 0 & S_{22} \end{bmatrix}$$

উপরে আলোচিত উপাত্তের ক্ষেত্রে ধরা যাক Σ-এর মান

$$\Sigma_0 = \begin{bmatrix} 1.80 & 1.00 \\ 1.00 & 7.00 \end{bmatrix}, \text{ এখানে } p = 2$$

তাহলে $\hat{k} = \text{tr}\,\Sigma_0^{-1}S/p$

$$= \text{tr}\begin{bmatrix} 0.6034 & -0.0862 \\ -0.0862 & 0.1552 \end{bmatrix}\begin{bmatrix} 1.7076 & 1.0711 \\ 1.0711 & 7.2622 \end{bmatrix}\Big/2$$

$$= \text{tr}\begin{bmatrix} 0.9380 & 0.0203 \\ 0.0190 & 1.0348 \end{bmatrix}\Big/2 = 0.9864$$

এখন $\hat{\Sigma} = \hat{k}\Sigma_0 = \begin{bmatrix} 1.7755 & 0.9864 \\ 0.9864 & 6.9048 \end{bmatrix}$

k নমুনাভিত্তিক নিরূপণ (Estimation on the Basis of k Samples) : ধরা যাক $X_1, X_2, \ldots, X_k$ হলো k অনপেক্ষ উপাত্ত ম্যাট্রিক্স যেখানে $X_i(n_i \times p)$ অনপেক্ষভাবে একই $N_p(\mu_i, \Sigma_i)$, $i = 1, 2, \ldots, k$ বিন্যাস অনুসরণ করে। এখানে μ_i ও Σ_i নিরূপণ করার জন্য দুটি শর্ত বিবেচনা করা যেতে পারে। যেমন :

(i) $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$

আবার (ii) $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$ এবং $\mu_1 = \mu_2 = \ldots = \mu_k$

অবশ্য দ্বিতীয় শর্তের ক্ষেত্রে সকল উপাত্ত ম্যাট্রিক্সকে একটি $(n \times p)$ আকারের উপাত্ত ম্যাট্রিক্স বিবেচনা করা যেতে পারে, যেখানে $n = \Sigma n_i$ ।

প্রথম শর্তের ক্ষেত্রে সম্ভাব্যতা ফাংশন (LF)-কে লেখা যায়

$$\log L = -\frac{1}{2}\sum_i [n_i \log |2\pi\Sigma| + n_i \operatorname{tr} \Sigma^{-1}\{S_i + (\bar{X}_i - \mu)(\bar{X}_i - \mu)'\}]$$

এখানে S_i হলো i-তম নমুনা হতে প্রাপ্ত সহ-ভেদাঙ্ক ম্যাট্রিক্স। যেহেতু μ_i এর উপর কোনো শর্ত আরোপ করা হয় নি, সে কারণে $\hat{\mu}_i = \bar{X}_i$ । আবার $n = \Sigma n_i$ এবং $W = \Sigma n_i S_i$ বিবেচনা করা হলে

$$\log L = -\frac{n}{2}\log |2\pi\Sigma| - \frac{1}{2}\operatorname{tr} \Sigma^{-1} W$$

এখন $\dfrac{\partial \log L}{\partial \Sigma} = 0$ হতে পাওয়া যায়

$$\hat{\Sigma} = n^{-1} W$$

এই $\hat{\Sigma}$ হলো প্রথম শর্তের ভিত্তিতে সর্বোচ্চ সম্ভাব্য নিরূপক [MLE]।

৩.২ যাচাই পদ্ধতি (Method of Test)

বহুচলক বিশ্লেষণের জন্য একটি মৌলিক অনুমান হলো যে, উপাত্ত ম্যাট্রিক্স বহুচলক পরিমিত বিন্যাস অনুসরণ করবে। ধরা যাক X হলো $(n \times p)$ অর্ডারের উপাত্ত ম্যাট্রিক্স যা $N_p(\mu, \Sigma)$ হতে চয়ন করা নমুনা। যেহেতু X, $N_p(\mu, \Sigma)$ অনুসরণ করে, X এর বিন্যাসের $\frac{1}{2}p(p+3)$ পরামান আছে এবং সে কারণে অন্যান্য নাস্তিকল্পনার কথা বাদ দিলেও এই পরামান ভেক্টরের ক্ষেত্রে $2^{p(p+3)/2}$ নাস্তিকল্পনা বিবেচনা করা যেতে পারে যা পরামান ভেক্টরের একটি উপ-গুচ্ছ এর মান নির্দিষ্ট করতে পারে। এছাড়া দুটি পরামানের অনুপাত-এর মান নির্দিষ্ট করার জন্য বা পরামানসমূহের ক্ষেত্রে কোনো রৈখিক বা অরৈখিক শর্ত আরোপ করা যায় কিনা সে সম্পর্কে নাস্তিকল্পনা বিবেচনা করতে হয়।

নাস্তিকল্পনা বিবেচিত হওয়ার পর ঐ নাস্তিকল্পনা যাচাই করার জন্য যাচাই তথ্যজমান নির্বাচিত করতে হয়, নাস্তিকল্পনার ভিত্তিতে যাচাই তথ্যজমান সঠিক কিনা সেটিও বিবেচনা করতে হয়। বর্তমান অনুচ্ছেদে মূলত একটি এবং দুটি নমুনার ভিত্তিতে μ এবং Σ এর জন্য নাস্তিকল্পনা যাচাই পদ্ধতি আলোচনা করা হবে। যাচাই পদ্ধতি হিসেবে (১) সম্ভাব্যতা অনুপাত যাচাই (likelihood ratio test, LRT) এবং (২) ইউনিয়ন ইন্টারসেকশন যাচাই (Union intersection test, UIT) পদ্ধতি বিশেষভাবে উল্লেখযোগ্য।

৩.২.১ μ এর জন্য Hotelling T^2-যাচাই (Hotelling T^2 test for μ)

২.৩ অনুচ্ছেদে Hotelling T^2 এর বিন্যাস আলোচনা করতে গিয়ে T^2 তথ্যজমানের উল্লেখ করা হয়েছে। তথ্যজমান ২.৪.১-এর ব্যুৎপত্তি হয়েছে সম্ভাব্যতা অনুপাত যাচাই (LRT)-এর ভিত্তিতে। এই তথ্যজমানটি Σ-এর মান জানা বিবেচনা করে $H_0 : \mu = \mu_0$ যাচাই করার জন্য ব্যবহৃত হয়। এখানে

$$-2\log\lambda = 2[\max \log L(\mu, \Sigma^{-1}) - \max \log L(\mu_0, \Sigma^{-1})]$$

$$= n(\bar{X} - \mu_0)' \Sigma^{-1} (\bar{X} - \mu_0) \qquad (৩.২.১)$$

এখানে X $(n \times p)$ হলো $N_p(\mu, \Sigma)$ হতে চয়ন করা নমুনা। $\bar{X}$ হলো নমুনা গড় ভেক্টর, $-2\log\lambda$ এর বিন্যাস হলো p স্বাধীনতার মাত্রাবিশিষ্ট কাইবর্গ (χ^2_p) বিন্যাস।

উদাহরণ হিসেবে ১.২ উদাহরণের উপাত্তের ক্ষেত্রে M. Tenelus এর দৈহিক দৈর্ঘ্য ও দৈহিক ওজনের গড় ভেক্টরের ভিত্তিতে নাস্তিকল্পনা $H_0 : \mu_0 = [59.12 \quad 3.63]'$ যাচাই করা যেতে পারে। এই μ_0 Bhuyan and Nair (1995) এর কাজ হতে উদ্ধৃত করা হয়েছে। তারা আরো দেখিয়েছেন যে

$$\Sigma = \begin{bmatrix} 135.71 & 21.99 \\ 21.99 & 4.31 \end{bmatrix}$$

কাজেই Σ-এর মান জানা বিবেচনা করে $H_0 : \mu = \mu_0$ যাচাই করা যায়। এখানে $\bar{X} = [46.80 \quad 1.94]'$। এখন

$$-2\log\lambda = n(\bar{X} - \mu_0)' \Sigma^{-1} (\bar{X} - \mu_0)$$

$$= 15 \left(\begin{bmatrix} 46.80 \\ 1.94 \end{bmatrix} - \begin{bmatrix} 59.12 \\ 3.63 \end{bmatrix} \right)' \begin{bmatrix} 0.0425 & -0.2170 \\ -0.2170 & 1.3390 \end{bmatrix}$$

$$\times \left(\begin{bmatrix} 46.80 \\ 1.94 \end{bmatrix} - \begin{bmatrix} 59.12 \\ 3.63 \end{bmatrix} \right)$$

$$= 39.39$$

দেখা যাচ্ছে যে $p[-2\log\lambda \geq 39.39] < 0.001$। কাজেই নাস্তিকল্পনা বাতিল। এখানে $-2\log\lambda$ এর বিন্যাস হলো 1 স্বাধীনতার মাত্রাবিশিষ্ট কাইবর্গ বিন্যাস।

বাস্তবে Hotelling T^2-এর ক্ষেত্রে Σ-এর মান জানা বিবেচনা করা হয় না। Σ-কে $H_0 : \mu = \mu_0$ এবং $H_0 : \mu \neq \mu_0$ এর ভিত্তিতে নিরূপণ করতে হয়। ৩.১ অনুচ্ছেদে লক্ষ্য করা গিয়েছে যে H_0 এর অধীনে

$$\hat{\mu} = \mu_0 \quad \text{এবং} \quad \hat{\Sigma} = S + (\bar{X} - \mu_0)(\bar{X} - \mu_0)'$$

এবং H_0-এর অধীনে

$$\hat{\mu} = \bar{X}, \quad \hat{\Sigma} = S$$

কাজেই $$\max \log L\left(\mu_0, \hat{\sum}_{H_0}\right) = -\frac{n}{2}\left\{p \log 2\pi + \log |S| + \log\left[1 + \left(X - \mu_0\right)' S^{-1}\left(\bar{X} - \mu_0\right)\right] + p\right\}$$

এবং $$\max \log L\left(\bar{X}, \hat{\sum}_{H_A}\right) = -\frac{n}{2}\left\{p \log 2\pi + \log |S| + p\right\}$$

$\therefore$ $$-2 \log \lambda = n \log\left[1 + (\bar{X} - \mu_0)' S^{-1} (\bar{X} - \mu_0)\right] \qquad (৩.২.২)$$

এখানে ৩.২.২ এর বিন্যাস হলো $T^2(p, n-1)$ এবং এটি একটি নমুনার জন্য Hotelling T^2-তথ্যজমান। একে লেখা যায়

$$T^2 = (n-1)(\bar{X} - \mu_0)' S^{-1} (\bar{X} - \mu_0) \qquad (৩.২.৩)$$

আবার, $\frac{n-p}{p}(\bar{X} - \mu_0)' S^{-1}(\bar{X} - \mu_0)$ এর বিন্যাস হলো $F_{p, n-p}$।

উপরে আলোচিত উদাহরণের ক্ষেত্রে

$$S = \begin{bmatrix} 10.5600 & 1.3147 \\ 1.3147 & 0.2331 \end{bmatrix}, \quad p = 2, \quad n = 15$$

কাজেই $$F = \frac{n-p}{p}(\bar{X} - \mu_0)' S^{-1} (\bar{X} - \mu_0)$$

$$= \frac{15-2}{2}\begin{bmatrix} -12.32 & -1.69 \end{bmatrix}\begin{bmatrix} 0.3176 & -1.7914 \\ -1.7914 & 14.3891 \end{bmatrix}\begin{bmatrix} -12.32 \\ -1.69 \end{bmatrix}$$

$$= 95.59$$

এখন $p(F_{2,13} = 95.59) < 0.001$ হওয়াতে নাস্তিকল্পনা বাতিল বলে বিবেচিত হলো।

৩.২.২ μ এর জন্য ইউনিয়ন ইন্টারসেকশন যাচাই (Union Intersection Test for μ)

ধরা যাক X $(n \times p)$ অর্ডারের উপাত্ত ম্যাট্রিক্স এবং তা $N_p(\mu, \Sigma)$ হতে চয়ন করা নমুনা। ধরা যাক Σ এর সকল বস্তু অজানা। যাচাই করতে হবে $H_0 : \mu = \mu_0$। বিকল্প কল্পনা হলো $H_A : \mu \neq \mu_0$।

ধরা যাক n অনপেক্ষ উপাত্ত ভেক্টরের ভিত্তিতে μ ও Σ-এর নিরূপক, যথাক্রমে $\bar{X}$ ও S। এখন নাস্তিকল্পনা যাচাই করার জন্য X এর একটি রৈখিক সংযোগ (linear combination) $a'X$ বিবেচনা করা যাক, এখানে a হলো বাস্তব বস্তু-ভিত্তিক $(p \times 1)$ অর্ডারের একটি শূন্য নয় (non null) এমন ভেক্টর। সুতরাং $a'X \sim N(a'\mu, a'\Sigma a)$ এবং এটি একচলক পরিমিত বিন্যাস। এই $a'X$-এর ক্ষেত্রে একচলক নাস্তিকল্পনা হলো $H_0 : a'\mu = a'\mu_0$ এবং যাচাই তথ্যজমান হলো

$$t(a) = \frac{a'(\bar{X} - \mu_0)\sqrt{n}}{\sqrt{a'Sa}} \qquad (৩.২.৪)$$

এই তথ্যজমানের ভিত্তিতে গ্রহণীয় এলাকা (acceptance region) হলো

$$t^2(a) \leqslant t^2_{\alpha/2,\, n-1} \quad .$$

উল্লিখিত যাচাই তথ্যজমানের যে নাস্তিকল্পনা বিবেচনা করা হয়েছে তার প্রাসঙ্গিক বহুচলক নাস্তিকল্পনা হবে $H_0 : a'\mu = a'\mu_0$ যা শূন্য নয় এমন a এর সকল মানের জন্য প্রযোজ্য। উক্ত নাস্তিকল্পনা গ্রহণযোগ্য হতে হলে a এর সকল মানের জন্য প্রতিটি একচলক নাস্তিকল্পনা গ্রহণযোগ্য হতে হবে। কাজেই a এর বিভিন্ন মানের জন্য নাস্তিকল্পনা বিবেচনা করা হলে বহুচলক নাস্তিকল্পনার ক্ষেত্রে গ্রহণযোগ্য এলাকা হবে সকল একচলক নাস্তিকল্পনার জন্য গ্রহণযোগ্য এলাকার ইন্টারসেকশন। অর্থাৎ

$$\bigcap_a \left[t^2(a) \leqslant t^2_{\alpha/2,\, n-1}\right]$$

কিন্তু $t^2(a)$-এর সকল মান উক্ত এলাকায় অবস্থিত হওয়ার অর্থ হলো $\max t^2(a) \leqslant t^2_{\alpha/2,\, n-1}$ শর্ত পূরণ করা। কাজেই $\max t^2(a)$ এর গড় ভেক্টরই হবে বহু-চলক নাস্তিকল্পনা যাচাই-এর জন্য যাচাই তথ্যজমান। কাজেই যাচাই তথ্যজমান পাওয়ার জন্য $t^2(a)$-কে সর্বোত্তম (maximize) করতে হবে। এটি করার জন্য বিবেচনা করা যাক যে, $t^2(a)$ মাত্রাবিহীন (dimensionless) এবং a-এর বস্তু-সমূহের মাপনী (scale) পরিবর্তন দ্বারা ক্ষতিগ্রস্ত হয় না। এর জন্য একটি শর্ত $a'Sa = 1$ আরোপ করা যাক। উক্ত শর্তের অধীনে

$$t^2(a) = a'(\bar{X} - \mu_0)(\bar{X} - \mu_0)'a\, n$$

এখন $t^2(a)$-কে সর্বোত্তম করার জন্য Lagranges গুণনীয়ক (Multiplier) λ ব্যবহার করে লেখা যায়

$$t^2(a) = a'\left[(\bar{X} - \mu_0)(\bar{X} - \mu_0)'n - \lambda S\right]a \qquad (৩.২.৫)$$

এখন ৩.২.৫-কে a এর প্রসঙ্গে বিযোজন (differentiation) করে পাওয়া যায়

$$[(\bar{X}-\mu_0)(\bar{X}-\mu_0)'n-\lambda S]a=0$$

উভয় পাশে a দ্বারা প্রাক গুণ করে পাওয়া যায়

$$a'[(\bar{X}-\mu_0)(\bar{X}-\mu_0)'n-\lambda S]a=0$$

এখন সমাধান করে পাওয়া যায়

$$\lambda=\frac{a'(\bar{X}-\mu_0)(\bar{X}-\mu_0)'an}{a'Sa}$$

$$=\frac{[a'(\bar{X}-\mu_0)]^2n}{a'Sa}=t^2(a)$$

এখন λ-কে লেখা যায়

$$\lambda=S^{-1}(\bar{X}-\mu_0)(\bar{X}-\mu_0)'n$$

$$=\text{tr}\,S^{-1}(\bar{X}-\mu_0)(\bar{X}-\mu_0)'n$$

$$=n(\bar{X}-\mu_0)'S^{-1}(\bar{X}-\mu_0)$$

$$=T^2$$

নাস্তিকল্পনার অধীনে T^2-কে F-এর মাধ্যমে প্রকাশ করা যায়, যেখানে

$$F=\frac{n-p}{p(n-1)}T^2$$

এই F হলো p এবং $(n-p)$ স্বাধীনতার মাত্রাবিশিষ্ট। এখানে T^2 হলো এক নমুনার জন্য Hotelling T^2।

উপরে আলোচিত যাচাই পদ্ধতির ক্ষেত্রে Σ-এর মান জানা থাকলে $H_0: a'\mu=a'\mu_0$ এর জন্য পাওয়া যায়

$$Z^2_{(a)}=n\,a'(\bar{X}-\mu_0)(\bar{X}-\mu_0)'a/a'\Sigma a \qquad (৩.২.৬)$$

নাস্তিকল্পনা বাতিল হবে

$$\max Z^2_a=n(\bar{X}-\mu_0)'\Sigma^{-1}(\bar{X}-\mu_0) \qquad (৩.২.৭)$$

এর জন্য। এখানে ৩.২.৭ এর বিন্যাস হলো χ^2_p। দেখা যাচ্ছে যে, $H_0: a'\mu=a'\mu_0$ যাচাই করার জন্য UIT এবং LRT একই যাচাই তথ্যজমান সরবরাহ করে।

৩.২.৩ দুই নমুনার জন্য Hotelling T^2 যাচাই (Hotelling T^2 test for two Samples)

ধরা যাক X_1 $(n_1 \times p)$ ও X_2 $(n_2 \times p)$ হলো দুটি উপাত্ত ম্যাট্রিক্স যেখানে $X_i (i = 1, 2)$ হলো $N_p(\mu_i, \Sigma_i)$ হতে চয়ন করা নমুনা। যাচাই করতে হবে $H_0 : \mu_1 = \mu_2$ । ২.১১ উপপাদ্যের মাধ্যমে প্রমাণ করা হয়েছে যে $\Sigma_1 = \Sigma_2$ হলে $H_0 : \mu_1 = \mu_2$ এর জন্য যাচাই তথ্যজমান হলো

$$\frac{n_1 n_2}{n} D^2 = \frac{n_1 n_2}{n} (\bar{x}_1 - \bar{x}_2)' S_S^{-1} (\bar{x}_1 - \bar{x}_2) \qquad (৩.২.৭)$$

যেখানে $\frac{n_1 n_2}{n} D^2 \sim T^2 (p, n-2)$, $\bar{x}_i$ হলো i-তম নমুনা গড় ভেক্টর, $S_S = (n+2)^{-1} (n_1 S_2 + n_2 S_2)$ । এছাড়া আরো দেখানো হয়েছে যে,

$$\frac{n_1 n_2 (n-p-1) D^2}{n(n-2)p} \sim F_{p, n-p-1}$$

এখানে $n = n_1 + n_2$ ।

উদাহরণ হিসেবে ১.২ উদাহরণের উপাত্তের ক্ষেত্রে M. Rusticus ও M. Tenellus এর দৈহিক দৈর্ঘ্য ও দৈহিক ওজনের মধ্যে পার্থক্য আছে কিনা তা যাচাই করা যেতে পারে। উক্ত উদাহরণের ক্ষেত্রে M. Rusticus এর জন্য

$$\bar{X}_1 = \begin{bmatrix} 38.80 & 2.68 \end{bmatrix}', \quad S_1 = \begin{bmatrix} 3.8933 & 0.5960 \\ 0.5960 & 1.2896 \end{bmatrix}$$

এবং M. Tenellus এর জন্য

$$\bar{X}_2 = \begin{bmatrix} 46.80 & 1.94 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 10.5600 & 1.3147 \\ 1.3147 & 0.2331 \end{bmatrix}$$

যাচাই করতে হবে $H_0 : \mu_1 = \mu_2$ । এখানে অনুমান হলো M. Rusticus এর উপাত্ত $N_2(\mu_1, \Sigma_1)$ হতে এবং M. Tenellus এর উপাত্ত $N_2(\mu_2, \Sigma_2)$ হতে চয়ন করা হয়েছে এবং $\Sigma_1 = \Sigma_2$ ।

এখন উপরিউক্ত উপাত্ত হতে পাওয়া যায়

$$S_S = (30+2)^{-1} \left(\begin{bmatrix} 58.3995 & 8.9400 \\ 8.9400 & 19.3440 \end{bmatrix} + \begin{bmatrix} 158.4000 & 19.7205 \\ 19.7205 & 3.4965 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 6.7750 & 0.8956 \\ 0.8956 & 0.7138 \end{bmatrix}$$

তাহলে

$$\frac{n_1 n_2}{n} D^2 = \left[[38.80 \;\; 2.68] - [46.80 \;\; 1.94] \right] \begin{bmatrix} 0.1770 & -0.2220 \\ -0.2220 & 1.6795 \end{bmatrix} \times \left[\begin{bmatrix} 38.80 \\ 2.68 \end{bmatrix} - \begin{bmatrix} 46.80 \\ 1.94 \end{bmatrix} \right] \frac{15 \times 15}{30}$$

$$= [-8.00 \;\; 0.74] \begin{bmatrix} 0.1770 & -0.2220 \\ -0.2220 & 1.6795 \end{bmatrix} \begin{bmatrix} -8.00 \\ 0.74 \end{bmatrix} 7.5$$

$$= 111.57$$

$$\therefore \quad F = \frac{n_1 n_2 (n - p - 1) D^2}{n(n-2)p} = 53.79$$

এখন $p(F_{2, 12} > 53.79) < 0.01$ হওয়াতে নাস্তিকল্পনা বাতিল বলে বিবেচিত হলো। দৈহিক ওজন এবং দৈহিক দৈর্ঘ্য এর ভিত্তিতে বলা যায় যে M. Rusticus ও M. Tenellus দুটি ভিন্ন slugs।

৩.২.৪ বহু নমুনার ক্ষেত্রে গড় ভেক্টরের সমতা যাচাই (Test of Equality of Mean Vector for Multi-Sample)

ধরা যাক X_i $(n_i \times p)$ হলো i-তম উপাত্ত ম্যাট্রিক্স যা $N_p(\mu_i, \Sigma_i)$ হতে চয়ন করা নমুনা $(i = 1, 2, \ldots, k)$। যাচাই করতে হবে $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$। অনুমান করা যাক যে $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$।

উপরিউক্ত নাস্তিকল্পনা যাচাই করার জন্য LRT এবং UIT উভয় পদ্ধতিই প্রয়োগ করা যাক।

LRT (Wilk's $\wedge$ যাচাই) : নাস্তিকল্পনা এবং অনুমানের ভিত্তিতে k নমুনাকে একটি নমুনা বিবেচনা করা যায়। উক্ত নমুনার ভিত্তিতে μ ও Σ-এর সর্বোচ্চ সম্ভাব্যতা নিরূপক (MLE) হলো যথাক্রমে $\bar{X}$ ও S। আবার বিকল্প কল্পনার অধীনে μ_i এর MLE হলো $\bar{X}_i$ এবং সাধারণ সহ-ভেদাঙ্ক ম্যাট্রিক্স-এর MLE হলো $n^{-1}W \left[\sum^{k} n_i = n \right]$, যেখানে $W = \sum^{k} n_i S_i$। তাহলে LRT নির্দেশক হবে—

$$\lambda = \{ |W| / |nS| \}^{n/2} = |T^{-1} W|^{n/2} \qquad (৩.২.৮)$$

এখানে $T = nS$, $B = T - W = \Sigma n_i (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})'$। এখন

$$\lambda^{2/n} = |W| / |B + W| = |I + W^{-1} B|^{-1} \qquad (৩.২.৯)$$

নাস্তিকল্পনা ও অনুমানের ভিত্তিতে উপাত্ত ম্যাট্রিক্সকে লেখা যায়

$$\mathbf{X}\ (n \times p) = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}$$

ধরা যাক $\mathbf{1}_i$ হলো $(n \times 1)$ অর্ডারের এমন একটি ভেক্টর যার বস্তুগুলো হলো i-তম নমুনা অনুযায়ী স্থানে 1 এবং অন্যস্থানে শূন্য। আরো ধরা যাক $I_i = \text{diag}(\mathbf{1}_i)$। তাহলে $I = \Sigma I_i$ এবং $\mathbf{1} = \Sigma \mathbf{1}_i$। নমুনা সহ-ভেদাঙ্ক ম্যাট্রিক্সকে লেখা যায় $n_i S_i = X' H_i X$, যেখানে $H_i = I_i - n_i^{-1} \mathbf{1}_i \mathbf{1}_i'$। এখন $C_1 = \Sigma H_i$ এবং $C_2 = \Sigma n_i^{-1} \mathbf{1}_i \mathbf{1}_i' - n^{-1} \mathbf{1}\mathbf{1}'$ ধরা হলে $W = X' C_1 X$ এবং $B = X' C_2 X$ লেখা যায়। এখানে C_1 ও C_2 হলো যথাক্রমে $n-k$ ও $k-1$ পদসংখ্যা (Rank) বিশিষ্ট আইডেমপোটেন্ট ম্যাট্রিক্স, তাছাড়া $C_1 C_2 = 0$। এখন অনুসিদ্ধান্ত (২.১৩) ও (২.১৫) অনুসারে লেখা যায়

$$W = X' C_1 X \sim W_p(n-k, \Sigma)$$

এবং

$$B = X' C_2 X \sim W_p(k-1, \Sigma)$$

এছাড়া W ও B হলো অনপেক্ষ। কাজেই $n \geqslant p+k$ হলে

$$|\ I + W^{-1} B\ |^{-1} \sim \Lambda\,(p,\ n-k,\ k-1)$$

আবার, $$\frac{(n-k-p+1)\ \{1 - \Lambda^{1/2}(p, n-k, k-1)\}}{p\,\Lambda^{1/2}(p,\ n-k,\ k-1)} \sim F_{2p,\ 2(n-k-p+1)}$$

উদাহরণ হিসেবে ১.২ উদাহরণের উপাত্তের ক্ষেত্রে 4 প্রকার slugs এর গড় দৈহিক দৈর্ঘ্য ও গড় দৈহিক ওজনবিশিষ্ট 4 ভেক্টরের মধ্যে পার্থক্য আছে কিনা যাচাই করে দেখা যেতে পারে। ঐ উদাহরণের ক্ষেত্রে M. Rusticus, M. Gagates, M. Tenellus এবং M. Sowerbye এর জন্য গড় ভেক্টর হলো যথাক্রমে $\bar{X}_1 = [38.80\ \ 2.68]'$, $\bar{X}_2 =]41.07\ \ 3.23]'$, $\bar{X}_3 = [46.80\ \ 1.94]'$ $\bar{X}_4 = [49.40\ \ 2.75]'$। এগুলোর প্রাসঙ্গিক সহ-ভেদাঙ্ক ম্যাট্রিক্স হলো

$$S_1 = \begin{bmatrix} 3.8933 & 0.5960 \\ 0.5960 & 1.2896 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 7.2622 & -3.4622 \\ -3.4622 & 1.7076 \end{bmatrix}$$

$$S_3 = \begin{bmatrix} 10.5600 & 1.3147 \\ 1.3147 & 0.2331 \end{bmatrix}, \quad S_4 = \begin{bmatrix} 59.8400 & 6.8280 \\ 6.8280 & 1.0118 \end{bmatrix}$$

$$W = \Sigma n_i S_i = \begin{bmatrix} 1223.3325 & 79.1475 \\ 79.1475 & 63.6315 \end{bmatrix}$$

$$\bar{X} = [\, 44.02 \quad 2.65 \,]'$$

$$B = \Sigma n_i (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})'$$

$$= 15\left[\left\{ \begin{bmatrix} 38.80 \\ 2.68 \end{bmatrix} - \begin{bmatrix} 44.02 \\ 2.65 \end{bmatrix} \right\} \left\{ \begin{bmatrix} 38.80 \\ 2.68 \end{bmatrix} - \begin{bmatrix} 44.02 \\ 2.65 \end{bmatrix} \right\}' \right.$$

$$+ \left\{ \begin{bmatrix} 41.07 \\ 3.23 \end{bmatrix} - \begin{bmatrix} 44.02 \\ 2.65 \end{bmatrix} \right\} \left\{ \begin{bmatrix} 41.07 \\ 3.23 \end{bmatrix} - \begin{bmatrix} 44.02 \\ 2.65 \end{bmatrix} \right\}'$$

$$+ \left\{ \begin{bmatrix} 46.80 \\ 1.94 \end{bmatrix} - \begin{bmatrix} 44.02 \\ 2.65 \end{bmatrix} \right\} \left\{ \begin{bmatrix} 46.80 \\ 1.94 \end{bmatrix} - \begin{bmatrix} 44.02 \\ 2.65 \end{bmatrix} \right\}'$$

$$\left. + \left\{ \begin{bmatrix} 49.40 \\ 2.75 \end{bmatrix} - \begin{bmatrix} 44.02 \\ 2.65 \end{bmatrix} \right\} \left\{ \begin{bmatrix} 49.40 \\ 2.75 \end{bmatrix} - \begin{bmatrix} 44.02 \\ 2.65 \end{bmatrix} \right\}' \right]$$

$$= \begin{bmatrix} 1089.3555 & -49.5510 \\ -49.5510 & 12.7710 \end{bmatrix}$$

$$B + W = \begin{bmatrix} 2312.6880 & 29.5965 \\ 29.5965 & 76.4025 \end{bmatrix}$$

এখন $\lambda^{2/n} = \Lambda(p, n-k, k-1) = |W| / |B+W| = 0.4071$

আবার $$F_{2p,\, 2(n-k-p+1)} = \frac{(15-4-2+1)\{1-\sqrt{0.4071}\}}{2\sqrt{0.4071}}$$

$$F_{4,20} = 2.84$$

$p(F_{4,20} \geq 2.84) > 0.05$ হওয়াতে 4 প্রকার slugs তাদের দৈহিক দৈর্ঘ্য ও দৈহিক ওজনের ভিত্তিতে একই প্রকার বিবেচিত হতে পারে।

UIT : এতক্ষণ $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ যাচাই করার জন্য LRT পদ্ধতি আলোচনা করা হয়েছে। এখন উক্ত নাস্তিকল্পনার জন্য UIT পদ্ধতি আলোচনা করা হবে। এই পদ্ধতি একচলক নাস্তিকল্পনার সাদৃশ্য। একচলকের ক্ষেত্রে যাচাই তথ্যজমান হলো

$$\Sigma n_i (\bar{x}_i - \bar{x}) / \Sigma n_i s_i^2 \qquad (৩.২.১০)$$

এখানে $\bar{x}$ হলো সার্বিক গড়। এখন a যদি $(n \times 1)$ বাস্তব মানের ভেক্টর হয়, তাহলে Xa রৈখিক সংযোগের ক্ষেত্রে (৩.২.১০) এর প্রাসঙ্গিক সূত্র হবে

$$\Sigma n_i \{a' (\bar{X}_i - \bar{X})\}^2 / \Sigma n_i a' S_i a \qquad (৩.২.১১)$$

উক্ত সূত্রের সর্বোচ্চ মান হবে

$$W^{-1}B=(\Sigma n_iS_i)^{-1}\Sigma n_i(\bar{X}_i-\bar{X})(\bar{X}_i-\bar{X})'$$

এর সবচেয়ে বড় আইগেন মান (Eigen value)। ধরা যাক U_a হলো যাচাই তথ্যজমান। তাহলে নাস্তিকল্পনা বাতিল হবে যদি $U_a \leqslant C_{1a}$ বা $U_a \geqslant C_{2a}$ হয়। এখানে C_{1a} এবং C_{2a} এমনভাবে নির্ণয় করতে হবে যেন যাচাই আকার (size of test) α এর সমান হয়। এ ক্ষেত্রে বর্জন ক্ষেত্র হবে

$$\lambda_p(W^{-1}B)<C_1 \text{ বা } \lambda_1(W^{-1}B)>C_2$$

এখানে λ_i হলো $W^{-1}B$ এর i-তম সর্বোচ্চ আইগেন মান।

৩.২.৫ এক নমুনার ক্ষেত্রে μ-এর কনট্রাস্ট যাচাই (Test of Contrast of μ in one Sample)

ধরা যাক X $(n\times p)$ হলো উপাত্ত ম্যাট্রিক্স যা $N_p(\mu, \Sigma)$ হতে চয়ন করা নমুনা। পরামান ভেক্টর μ এর একটি কনট্রাস্ট $R\mu$ বিবেচনা করা যাক। যাচাই করতে হবে $H_0 : R\mu=r$, যেখানে R এবং r পূর্ব নির্ধারিত মানবিশিষ্ট যথাক্রমে ম্যাট্রিক্স ও ভেক্টর। উক্ত যাচাই-এর জন্য Σ-এর মান সম্পর্কে অনুমান গুরুত্বপূর্ণ। এখানে Σ-এর মান জানা এবং অজানা উভয় প্রকার অনুমানের ভিত্তিতে যাচাই পদ্ধতি আলোচনা করা হবে। প্রথমে Σ-এর মান জানা বিবেচনা করা যাক।

$H_0 : R\mu=r$, Σ জানা : নাস্তিকল্পনার অধীনে μ এর MLE হলো

$$\hat{\mu}=\bar{X}-\Sigma R'\lambda=\bar{X}-\Sigma R'(R\Sigma R')^{-1}(R\bar{X}-r)$$

সুতরাং, LRT নির্দেশক অনুযায়ী যাচাই তথ্যজমান হলো

$$-2\log\lambda=n(R\bar{X}-r)'(R\Sigma R')^{-1}(R\bar{X}-r) \qquad (৩.২.১২)$$

এই যাচাই তথ্যজমানের বিন্যাস হলো χ^2_q, $q<p$। কারণ নাস্তিকল্পনার অধীনে XR' এর সারিগুলো অনপেক্ষভাবে একই $N_q(r, R\Sigma R')$ বিন্যাস অনুসরণ করে। এখানে q হলো ভেক্টর r এর বস্তুসমূহ (elements)। কাজেই উপপাদ্য ২.৭ অনুসারে $-2\log\lambda$ এর বিন্যাস হবে χ^2_q।

নাস্তিকল্পনা $H_0 : R\mu=r$ যাচাই করার সময় μ-কে দুটি ভাগে ভাগ করে যে কোনো এক ভাগের μ-গুলোর একটি নির্দিষ্ট মান সম্পর্কে নাস্তিকল্পনা যাচাই করা যায়। যেমন, ধরা যাক $\mu=[\mu_1' \mu_2']$ যাচাই করতে হবে $\mu_1=0$। এক্ষেত্রে নাস্তিকল্পনাকে লেখা যায়

$$H_0 : \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = r \qquad (৩.২.১৩)$$

এখানে $R = [\,1 \quad 0\,]$, $r = 0$। এই নাস্তিকল্পনার জন্য যাচাই তথ্যজমান হবে

$$-2\log\lambda = n\,\bar{X}_1'\,\Sigma_{11}^{-1}\,\bar{X}_1 \qquad (৩.২.১৪)$$

এখানে $\bar{X} = [\,\bar{X}_1' \quad \bar{X}_2'\,]'$, $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$

অনেক সময় $H_0 : \mu = k\mu_0$ নাস্তিকল্পনাটি যাচাই করার প্রয়োজন হয়। এখানে μ_0 হলো μ এর নির্দিষ্ট মান এবং k যে কোনো মান। এই নাস্তিকল্পনাকেও $H_0 : R\mu = 0$ হিসেবে লেখা যায়। সেক্ষেত্রে R ম্যাট্রিক্স হবে $(p-1)$ পদসংখ্যাবিশিষ্ট $(p-1)\times p$ অর্ডারের ম্যাট্রিক্স যার সারিগুলো μ_0 এর সমকৌণিক (orthogonal) হবে। নাস্তিকল্পনার অধীনে k এর MLE হলো

$$\hat{k} = \mu_0'\,\Sigma^{-1}\,\bar{X} \,/\, \mu_0'\,\Sigma^{-1}\,\mu_0 \qquad (৩.২.১৫)$$

এই $\hat{k}$ এর মান ব্যবহার করে নাস্তিকল্পনার জন্য যাচাই তথ্যজমান হলো

$$-2\log\lambda = n\,\bar{X}'\,\Sigma^{-1}\{\Sigma - (\mu_0'\,\Sigma^{-1}\,\mu_0)^{-1}\,\mu_0\mu_0'\}\,\Sigma^{-1}\,\bar{X} \qquad (৩.২.১৬)$$

এখানে $-2\log\lambda$ এর বিন্যাস হলো χ^2_{p-1}।

$H : R\mu = r$, Σ অজানা : Σ অজানা হলে μ এর MLE হলো

$$\hat{\mu} = \bar{X} - SR'\,(\,RSR'\,)^{-1}\,(R\bar{X} - r) \qquad (৩.২.১৭)$$

এখন ৩.২.২ প্রয়োগ করে পাওয়া যায়

$$-2\log\lambda = n\log(1 + d'\,S^{-1}d) \qquad (৩.২.১৮)$$

এখানে $d = SR'(RSR')^{-1}\,(R\,\bar{X} - r)$

উক্ত যাচাই ৩.২.১৮ তথ্যজমান

$$(n-1)\,d'\,S^{-1}\,d = (n-1)(R\,\bar{X} - r)'\,(RSR')^{-1}\,(R\,\bar{X} - r) \qquad (৩.২.১৯)$$

এর উপর ভিত্তি করে নির্ণয় করা হয়েছে। কিন্তু ৩.২.১৯ এর বিন্যাস হলো $T^2(q, n-1)$। সুতরাং ৩.২.১৮ এর বিন্যাস হলো $T^2(q, n-1)$ যা

$$F_{q,\,n-q} = \frac{n-q}{(n-1)q}\,T^2$$

বিন্যাস অনুসরণ করে। কারণ, নাস্তিকল্পনার অধীনে

$$R\,\bar{X} \sim N_q(r,\, n^{-1}\,R\Sigma R') \quad \text{এবং} \quad nRSR' \sim W_q(n-1,\, R\Sigma R')$$

এবং উহারা অনপেক্ষ।

এখন $H_0 : R\mu = r$, Σ-অজানা যাচাই করার জন্য 3.3 উদাহরণের ক্ষেত্রে দিনে 2 বার দোহন করা গরুর উপাত্তের ভিত্তিতে গরুর প্রাথমিক ওজন (B) ও দুধ উৎপাদনকাল শেষ হওয়ার পর ওজন (C) এর মধ্যে কোনো পার্থক্য হয়েছে কিনা তা লক্ষ্য করা যেতে পারে। এখানে

$$\bar{X}_2 = \begin{bmatrix} 778.04 \\ 782.68 \end{bmatrix} \text{ এবং } S_2 = \begin{bmatrix} 3548.82 & 2286.51 \\ 2286.51 & 4006.22 \end{bmatrix}$$

প্রাথমিক ওজন ও দুধ উৎপাদনকাল শেষ হওয়ার পর ওজনের মধ্যে কোনো পার্থক্য নেই বিবেচনা করার জন্য নাস্তিকল্পনা হলো

$$H_0 : R\mu = 0, \quad \text{এখানে } R = [1 \quad -1]$$

এখন যাচাই তথ্যজমান হলো

$$(n-1)(R\bar{X} - r)'(RSR')^{-1}(R\bar{X} - r)$$

$$= 27(-4.64)\,[2982.02]^{-1}(-4.64) = 1733445.84$$

এটি $T^2(1, 27)$। কাজেই

$$F_{1,27} = \frac{28-1}{1(27)} 1733445.84 = 1733445.84$$

দেখা যাচ্ছে যে $P(F_{1,27} = 1733445.84) < 0.0000$। কাজেই নাস্তিকল্পনা বাতিল। অর্থাৎ গরুর প্রাথমিক ওজন ও দুধ উৎপাদনকাল শেষ হওয়ার পর ওজনের মধ্যে অতিশয় তাৎপর্যপূর্ণ পার্থক্য আছে।

৩.২.৬ অসমসত্ত্ব সহ-ভেদাঙ্ক ম্যাট্রিক্স-এর ক্ষেত্রে গড় ভেক্টরের সমতা যাচাই (Test of Equality of Mean Vectors in Presence of Heterogeneous Covariance Matrix)

ধরা যাক $\bar{X}_1$ এবং $\bar{X}_2$ হলো অনপেক্ষ উপাত্ত ম্যাট্রিক্স, যেখানে $X_i\ (n_i \times p)$ হলো $N_p(\mu_i, \Sigma_i)$ হতে চয়ন করা নমুনা $(i = 1, 2)$। যাচাই করতে হবে $H_0 : \mu_1 = \mu_2$। জানা আছে যে $\Sigma_1 \neq \Sigma_2$ [অথবা যাচাই-এর (৩.২.২৬) মাধ্যমে সিদ্ধান্ত গৃহীত]। এখানে $H_0 : \mu_1 = \mu_2 \rightarrow H_0 : \mu_1 - \mu_2 = \delta = 0$। বিকল্প কল্পনা হলো $H_A : \delta \neq 0$।

একচলক বিশ্লেষণের ক্ষেত্রে এরূপ নাস্তিকল্পনা যাচাই Fisher-Behrens সমস্যা নামে পরিচিত। Yao(1965) উক্ত Fisher-Behrens সমস্যার সমাধান অনুযায়ী বহুচলক বিশ্লেষণের ক্ষেত্রে এই নাস্তিকল্পনা যাচাই পদ্ধতির প্রস্তাব করেছেন। এই যাচাই-এর জন্য ধরা যাক

$$\bar{X}_i \sim N_p(\mu_i, \Gamma_i), \quad S_i \sim W_p(f_i, \Gamma_i), \quad i = 1, 2$$

এখানে $f_1 = n_1 - 1$, $\Gamma_1 = n_1^{-1}\Sigma_1$, $i = 1, 2$ । ধরা যাক

$$d = \bar{X}_1 - \bar{X}_2 , \quad U_i = S_i/f_1$$

$$U = U_1 + U_2 , \quad \Gamma = \Gamma_1 + \Gamma_2$$

U হলো Γ-এর নিখুঁতি নিরূপক। এখন সাম্ভিকল্পনার অধীনে $d \sim N_p(0, \Gamma)$ । ধরা যাক $fU \sim W_p(f, \Gamma)$, যেখানে d এবং f অনপেক্ষ এবং f-এর মান এমনভাবে নিতে হবে যেন p-মাত্রার সকল a ভেক্টরের ক্ষেত্রে $fa'Ua \sim a'\Gamma a\chi_f^2$ হয়। তাহলে পাওয়া যায়

$$\omega_a = t_a^2(f) = (a'd)^2/a'Ua \qquad (৩.২.২০)$$

এখানে $t_a(f)$ হলো f স্বাধীনতার মাত্রাবিশিষ্ট 't' তথ্যজমান। কাজেই (৩.২.১) সেকশান-এ আলোচিত ফলাফলের ভিত্তিতে বলা যায়

$$\max_a \omega_a = \omega_a^* = d'U^{-1}d \sim T^2(p, f) \qquad (৩.২.২১)$$

এখানে a এর সর্বোচ্চ মান হলো $a^* = U^{-1}d$। Welch (1947) দেখিয়েছেন যে fU এর বিন্যাস $W_p(f, \Gamma)$ সব সময় সত্য না হলেও $\omega_a \sim t_a^2(f_a)$, যেখানে

$$\frac{1}{f_a} = \frac{1}{f_1}\left(\frac{a'U_1a}{a'Ua}\right)^2 + \frac{1}{f_2}\left(\frac{a'U_2a}{a'Ua}\right)^2$$

এখানে a^* এর মান বসিয়ে বলা যায় যে $d'U^{-1}d$ এর বিন্যাস $T^2(p, f^*)$ এর কাছাকাছি। যেখানে

$$\frac{1}{f^*} = \frac{1}{f_1}\left(\frac{d'U^{-1}U_1U^{-1}d}{d'U^{-1}d}\right)^2 + \frac{1}{f_2}\left(\frac{d'U^{-1}U_2U^{-1}d}{d'U^{-1}d}\right)^2$$

উদাহরণ হিসেবে উদাহরণ ১.১ এর উপাত্তের ভিত্তিতে দিনে ২ বার এবং দিনে ৩ বার দোহন করা গরুর প্রাথমিক ওজন ও দুধ উৎপাদনকাল শেষ হওয়ার পর ওজনের গড় ভেক্টর দুটির সমতা যাচাই করে দেখা যেতে পারে। উপাত্ত হতে পাওয়া যায়

$$\bar{X}_2 = \begin{bmatrix} 778.04 \\ 782.68 \end{bmatrix}, \quad \bar{X}_3 = \begin{bmatrix} 776.25 \\ 769.82 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} 3548.82 & 2286.51 \\ 2286.51 & 4006.22 \end{bmatrix}, \quad S_3 = \begin{bmatrix} 3660.04 & 2847.54 \\ 2847.54 & 3854.43 \end{bmatrix}$$

অনুমান করা হচ্ছে যে $\Sigma_2 \neq \Sigma_3$ (অবশ্য BOX'S M – test প্রয়োগ করে Σ_2 ও Σ_3 এর সমতা যাচাই করা যেতে পারে)।

এখানে

$$d=\begin{bmatrix}1.79\\12.86\end{bmatrix},\ U_2=\begin{bmatrix}131.44 & 84.69\\84.69 & 148.38\end{bmatrix},\ U_3=\begin{bmatrix}135.56 & 105.46\\105.46 & 142.76\end{bmatrix}$$

$$U=\begin{bmatrix}267.00 & 190.15\\190.15 & 291.14\end{bmatrix}$$

$$T^2(p, f^*)=d'U^{-1}d=[1.79\ \ 12.86]\begin{bmatrix}0.0070 & -0.0046\\-0.0046 & 0.0064\end{bmatrix}\begin{bmatrix}1.79\\12.86\end{bmatrix}$$

$$=0.8691$$

এখানে $p=2$ এবং $f^*=-53$। সুতরাং

$$F_{p,f^*-p+1}=\frac{(f^*-p+1)\,T^2(p, f^*)}{pf^*}$$

$$F_{2,52}=0.43$$

কিন্তু $p[p(F_{2,52}=0.43)]>0.05$ হওয়াতে গরুর প্রাথমিক ওজন ও দুধ উৎপাদন কাল শেষ হওয়ার পর ওজনের গড় ভেক্টর দুটি সমান বিবেচনা করা যেতে পারে।

৩.২.৭ সহ-ভেদাঙ্ক ম্যাট্রিক্স এর যাচাই (Test of Covariance Matrix)

পূর্ববর্তী অনুচ্ছেদসমূহে গড় ভেক্টরের সমতা যাচাই আলোচনা করতে গিয়ে সহ-ভেদাঙ্ক ম্যাট্রিক্স সম্পর্কে অনুমান করার বিষয় উল্লেখ করা হয়েছে। অনুমান না করে সহ-ভেদাঙ্ক ম্যাট্রিক্সসমূহের সমতা বা অসমতা যাচাই করেও সিদ্ধান্ত নেয়া যায়। বর্তমান অনুচ্ছেদে সহ-ভেদাঙ্ক ম্যাট্রিক্স সম্পর্কে নাস্তিকল্পনা যাচাই পদ্ধতি আলোচনা করা হবে।

মনে করি Σ হলো একটি গণসমষ্টি সহ-ভেদাঙ্ক ম্যাট্রিক্স। যাচাই করতে হবে যে

$$H: \Sigma=\sigma^2\begin{bmatrix}1 & \rho & \rho & \cdots & \rho\\\rho & 1 & \rho & \cdots & \rho\\\cdots & & \cdots & & \cdots\\\rho & \rho & \rho & \cdots & 1\end{bmatrix} \qquad (৩.২.২২)$$

উক্ত নাস্তিকল্পনার বিরুদ্ধে বিকল্প কল্পনা হলো :

H_A : Σ হলো একটি সাধারণ ধনাত্মক ডেফিনিট ম্যাট্রিক্স, [Wilks (1946)] ।

ধরা যাক Σ ম্যাট্রিক্স এর নিঝুঁকি নিরূপক হলো S এবং এর স্বাধীনতার মাত্রা হলো $(n-1)$ । এক্ষেত্রে σ^2 ও $\sigma^2\rho$ এর নিরূপক হলো যথাক্রমে

$$s^2 = \frac{1}{p}\sum_{i=1}^{p} s_{ii}$$

এবং $$s^2_r = \frac{1}{p(p-1)}\sum_{i\neq j}\sum s_{ij}, \quad \text{এখানে } S=(s_{ij})$$

এখন Wilks এর জেনারালাইজ্‌ড্‌ সম্ভাব্যতা-অনুপাত যাচাই (generalized likelihood ratio test) তথ্যজমান হলো

$$L = \frac{|S|}{(s^2)^p(1-r)^{p-1}[1+(p-1)r]} \qquad (৩.২.২৩)$$

Box (1949, 1950) দেখিয়েছেন যে, যাচাই তথ্যজমান

$$\chi^2 = -\left[(n-1) - \frac{p(p+1)^2(2p-3}{6(p-1)(p^2+p-4)}\right]\log_e L \qquad (৩.২.২৪)$$

$\{\frac{1}{2}p(p+1)-2\}$ স্বাধীনতার মাত্রাসহ প্রায় কাইবর্গ বিন্যাস অনুসরণ করে যদি নাস্তিকল্পনা সত্য হয় এবং $(n-1)$ বড় হয় ।

উদাহরণ হিসেবে ৩.২.৬ অনুচ্ছেদে উপস্থাপিত S_2 এর প্রাসঙ্গিক গণসমষ্টি সহ-ভেদাঙ্ক ম্যাট্রিক্স Σ_2 সম্পর্কে নাস্তিকল্পনা যাচাই করতে পারি । উক্ত S_2 এর ক্ষেত্রে $s^2 = 3777.52$, $s^2_r = 2286.51$, $r = 0.6053$, $|S^2| = 8989225.68$, $L = 0.9942$ । কাজেই $\chi^2 = 0.148$ । এখানে $p(\chi^2 \geqslant 0.148) > 0.05$ (স্বাধীনতার মাত্রা 1) হওয়ায় Σ_2 ম্যাট্রিক্স এর ক্ষেত্রে ভেদাঙ্কগুলো সমান এবং সহ-ভেদাঙ্কগুলোও সমান বিবেচনা করা যায় ।

এতক্ষণ Σ ম্যাট্রিক্স-এর একটি বিশেষ রূপ সম্পর্কে নাস্তিকল্পনা যাচাই করা হয়েছে । এই নাস্তিকল্পনার জন্য বিবেচনা করা হয়েছে একটি নমুনা । বাস্তবে p-মাত্রার k নমুনার বিশ্লেষণ প্রয়োজন হয় । ধরা যাক p-মাত্রার k সংখ্যক বহুচলক পরিমিত বিন্যাস আছে । মনে করি i-তম $(i = 1, 2, \ldots, k)$ বিন্যাসের সহ-ভেদাঙ্ক ম্যাট্রিক্স হলো Σ_i । যাচাই করতে হবে

$$H_0 : \Sigma_1 = \Sigma_2 = \cdots = \Sigma_k \qquad (৩.২.২৫)$$

এখানে বিকল্প কল্পনা হলো Σ_i একটি জেনারালাইজড ধনাত্মক ডেফিনিট ম্যাট্রিক্স $(i=1, 2, \ldots, k)$ ।

মনে করি i-তম গণসমষ্টি হতে n_i আকারের নমুনা চয়ন করা হয়েছে এবং S_i হলো Σ_i এর নির্ঝুঁকি নিরূপক, যেখানে S_i এর স্বাধীনতার মাত্রা (n_i-1)। নাস্তিকল্পনা সত্য হলে Σ_i এর নির্ঝুঁকি নিরূপক হবে

$$S=\frac{1}{n-k}\sum_{i=1}^{k}(n_i-1)S_i\,,\quad n=\sum_{i=1}^{k}n_i$$

এখন নাস্তিকল্পনা ৩.২.২৫ যাচাই করার জন্য যাচাই তথ্যজমান হলো

$$M=(n-k)\log_e|S|-\sum_{i=1}^{k}(n_i-1)\log_e|S_i| \qquad (৩.২.২৬)$$

Box (1949) দেখিয়েছেন যে MC^{-1} এর বিন্যাস $\frac{1}{2}(k-1)\,p(p+1)$ স্বাধীনতার মাত্রাবিশিষ্ট কাইবর্গ বিন্যাসের কাছাকাছি, যদি n_i বড় হয়। এখানে

$$C^{-1}=1-\frac{2p^2+3p-1}{6(p+1)(k-1)}\left[\sum_{i=1}^{k}\frac{1}{n_i-1}-\frac{1}{n-k}\right]$$

যদি সকল n_i সমান হয়, তাহলে

$$C^{-1}=1-\frac{(2p^2+3p-1)(k+1)}{6(p+1)kn}$$

উপরিউক্ত যাচাই তথ্যজমান ৩.২.২৬ Box's M-test নামে পরিচিত। যদি p এবং k চার বা পাঁচের বেশি না হয় এবং যদি n_i 20 বা তার বেশি হয়, তাহলে MC^{-1} এর বিন্যাস ভালভাবেই কাইবর্গ বিন্যাস অনুসরণ করবে। আবার p এবং k এর মান বড় হলে এবং n_i ছোট হলে নাস্তিকল্পনা ৩.২.২৫ এর জন্য Box F-যাচাই তথ্যজমান-এর প্রস্তাব করেছেন [Pearson and Hartley(1972)]। অন্যদিকে Greenstreet and Connor (1974) MC^{-1} যাচাই তথ্যজমানের ভিত্তিতে যাচাই শক্তি নির্ণয় করেছেন। Roy, Pillai এবং অন্যান্যরা $k=2$ হলে বিকল্প যাচাই পদ্ধতির প্রস্তাব করেছেন।

উদাহরণ হিসেবে ১.১ উদাহরণের উপাত্তের ভিত্তিতে দিনে ২ বার এবং দিনে ৩ বার দোহন করা গরুর প্রাথমিক ওজন ও দুধ উৎপাদনকাল শেষ হওয়ার পর ওজনের ভিত্তিতে প্রাপ্ত সহ-ভেদাঙ্ক ম্যাট্রিক্স ব্যবহার করে নাস্তিকল্পনা ৩.২.২৫ যাচাই করতে পারি। উপাত্ত হতে পাওয়া গেছে

$$S_2 = \begin{bmatrix} 3548.82 & 2286.51 \\ 2286.51 & 4006.22 \end{bmatrix}, \quad S_3 = \begin{bmatrix} 3660.04 & 2847.54 \\ 2847.54 & 3854.43 \end{bmatrix}$$

$n_2 = 28, \quad n_3 = 28, \quad k = 2, \quad n = 56$

তাহলে $S = \begin{bmatrix} 3604.43 & 2567.03 \\ 2567.03 & 3930.33 \end{bmatrix}, \quad |S_2| = 8989225.68$

$$|S_3| = 5998883.92, \quad |S| = 7576956.34$$

$$M = 1.69, \quad C^{-1} = 0.9599, \quad MC^{-1} = 1.62$$

এই MC^{-1} এর বিন্যাস হলো 3 স্বাধীনতার মাত্রাবিশিষ্ট কাইবর্গ বিন্যাস। কিন্তু $P(\chi^2 \geq 1.62) > 0.05$ হওয়ায় নাস্তিকল্পনা $H_0 : \Sigma_2 = \Sigma_3$ সত্য বলে বিবেচনা করা যায়।

৩.২.৮ পূর্ণ সমতা যাচাই (Test of Complete Homogeneity)

নাস্তিকল্পনা $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ যাচাই করার সময় শর্তারোপ করা হয়েছে যে $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$ (৩.২.৪ অনুচ্ছেদ)। আবার, (৩.২.৭) অনুচ্ছেদে আলোচনা করা হয়েছে নাস্তিকল্পনা $H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$ যাচাই পদ্ধতি। প্রথমোক্ত নাস্তিকল্পনার জন্য সম্ভাব্যতা অনুপাত যাচাই [LRT] তথ্যজমান হলো

$$\lambda_1 = |I + W^{-1}B|^{-n/2} \qquad [\text{(৩.২.৯) সমীকরণ অনুযায়ী}]$$

শেষোক্ত নাস্তিকল্পনার জন্য সম্ভাব্যতা অনুপাত যাচাই তথ্যজমান হলো

$$-2\log\lambda_2 = (n-k)\log|S| - \sum_{i=1}^{k}(n_i-1)\log|S_i|$$

এখন যাচাই করতে হবে

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \text{ এবং } \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k \qquad (৩.২.২৭)$$

উক্ত নাস্তিকল্পনার জন্য সম্ভাব্যতা অনুপাত যাচাই তথ্যজমান হলো

$$-2\log\lambda_3 = -2\log\lambda_2 - 2\log\lambda_1 \qquad (৩.২.২৮)$$

যাচাই তথ্যজমান (৩.২.২৮) $\frac{1}{2}p(k-1)(p+3)$ স্বাধীনতার মাত্রাসহ অভিসারীভাবে কাইবর্গ বিন্যাস অনুসরণ করে।

যাচাই তথ্যজমান (৩.২.২৮)-কে লেখা যায়

$$-2\log\lambda_3 = (n-k)\log|W_1| - \sum_{i=1}^{k}(n_i-1)\log|S_i| \qquad (৩.২.২৯)$$

এখানে S_i হলো Σ_i-এর নিখুঁতি নিরূপক এবং $W_1 = \dfrac{\Sigma(n_i-1)S_i}{n-k}$

উদাহরণ হিসেবে ১.২ উদাহরণের উপাত্ত ব্যবহার করে M. Rusticus, M. Gagates, M. Tenelus এবং M. Sowerbye এর গড় ভেক্টরগুলো ও সহ-ভেদাঙ্ক ম্যাট্রিক্সগুলোর সমতা যাচাই করা যেতে পারে। ঐ উপাত্তের ক্ষেত্রে

$$\bar{X}_1 = [\,38.80 \quad 2.68\,]\,, \quad \bar{X}_2 = [\,41.07 \quad 3.23\,]'$$

$$\bar{X}_3 = [\,46.80 \quad 1.94\,]' \quad \text{এবং} \quad \bar{X}_4 = [\,49.40 \quad 2.75\,]'$$

এগুলোর প্রাসঙ্গিক সহ-ভেদাঙ্ক ম্যাট্রিক্সগুলো হলো

$$S_1 = \begin{bmatrix} 3.8933 & 0.5960 \\ 0.5960 & 1.2896 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 7.2622 & -3.4622 \\ -3.4622 & 1.7076 \end{bmatrix}$$

$$S_3 = \begin{bmatrix} 10.5600 & 1.3147 \\ 1.3147 & 0.2331 \end{bmatrix}, \quad S_4 = \begin{bmatrix} 59.8400 & 6.8280 \\ 6.8280 & 1.0118 \end{bmatrix}$$

তাহলে,

$$W_1 = \begin{bmatrix} 20.3889 & 1.3191 \\ 1.3191 & 1.0605 \end{bmatrix}, \; |W_1| = 19.8824\,, \; |S_1| = 4.6656$$

$$|S_2| = 0.4141\,, \quad |S_3| = 0.7331\,, \quad |S_4| = 13.9245$$

$-2\log\lambda_3 = 125.69$। এই $-2\log\lambda_3$ অভিসারীভাবে 15 স্বাধীনতার মাত্রাসহ কাইবর্গ বিন্যাস অনুসরণ করে। এখানে $P[\,\chi^2 \geq 125.69\,] < 0.001$ হওয়াতে গড় ভেক্টরসমূহ ও সহ-ভেদাঙ্ক ম্যাট্রিক্সগুলোর ভিত্তিতে slug গুলোকে একই জাতীর বলা যায় না।

———

চতুর্থ অধ্যায়

প্রধান উপাদান বিশ্লেষণ
(Principal Component Analysis)

৪.১ সূচনা (Introduction)

আগেই উল্লেখ করা হয়েছে যে বহুচলক বিশ্লেষণের ক্ষেত্রে একই বস্তুর অনেকগুলো বৈশিষ্ট্য পর্যালোচনা করা হয় বা পর্যালোচনার যোগ্য। অবশ্য সব বৈশিষ্ট্য বা চলকই যে বিশেষ বিশেষ উদ্দেশ্য সাধনের জন্য বিশ্লেষিত হওয়ার যোগ্য তা নয়। সেক্ষেত্রে গবেষক বিচার বিবেচনা করে কিছু চলকের বৈশিষ্ট্য পর্যালোচনা করে থাকে। প্রাক কমপিউটার যুগে এই বিচার বিবেচনার একটি শক্তিশালী গাণিতিক ভিত্তি ছিল না। যতই দিন যাচ্ছে কমপিউটারের আধুনিকায়ন হচ্ছে, যা গবেষকের গাণিতিক সমস্যা সমাধানের পথকে সহজতর করছে। ফলে অনাবশ্যক বিশ্লেষণ এড়িয়ে যথাযথ চলকের বৈশিষ্ট্য বিশ্লেষণ করে যে কোনো প্রাক-উদ্দেশ্যের প্রাসঙ্গিক উত্তর খুঁজে পাওয়া সহজ হচ্ছে। আবার, কোনো কোনো ক্ষেত্রে সব চলকের যুগপৎ বিশ্লেষণের পরিবর্তে সেগুলোর একটি রৈখিক সমাবেশ (linear combination)-এর বিশ্লেষণের ভিত্তিতেও গবেষকের প্রশ্নের অনুসন্ধিৎসার উত্তর পাওয়া যায়। প্রধান উপাদান বিশ্লেষণ এমন একটি পদ্ধতি যা মূল চলকসমূহকে অল্পসংখ্যক চলকে পরিবর্তন করে থাকে যেখানে ঐ অল্পসংখ্যক চলক হলো মূল চলকসমূহের রৈখিক সমাবেশ এবং এই রৈখিক সমাবেশ মূল চলক দ্বারা যে পরিমাণ ভেদাঙ্কের ব্যাখ্যা করা যায় তার বৃহত্তর অংশকেই ব্যাখ্যা করে থাকে। বর্তমান অধ্যায়ে প্রধান উপাদান বিশ্লেষণের বিভিন্ন দিক আলোচনা করা হবে।

৪.২ প্রধান উপাদান নির্ধারণ পদ্ধতি (Method of Extraction of Principal Component)

প্রধান উপাদান বিশ্লেষণ হলো মূল চলকগুচ্ছ ব্যবহার করে ঐ চলকগুচ্ছের কয়েকটি রৈখিক সমাবেশ নির্ণয় করা যেগুলো মূল উপাত্তসমূহকে ব্যাখ্যা করতে পারে। অবশ্য মূল উপাত্তের রৈখিক সমাবেশ উপাত্তের সম্পূর্ণ ভেদ ব্যাখ্যা করতে পারে না। তবে লক্ষ্য রাখতে হবে যে উপাত্ত বিশ্লেষণ করে যে পরিমাণ তথ্য পাওয়া যাওয়ার কথা তার পরিমাণ যেন খুব কম না হয়। এখানে রৈখিক সমাবেশ নির্ণয় করার ফলে সমাবেশকৃত চলকগুচ্ছের সংখ্যা মূল চলকগুচ্ছের সংখ্যার চেয়ে কম হয়। কম চলকগুচ্ছ দ্বারা মূল চলকের ভেদের বেশি অংশ ব্যাখ্যা করাই

প্রধান উপাদান বিশ্লেষণের মূল লক্ষ্য। অবশ্য রৈখিক সমাবেশের সংখ্যা মূল চলকের সংখ্যার চেয়ে কম হবেই এমন কোনো কথা নয়। প্রধান উপাদানসমূহের সংখ্যা মূল চলকের সংখ্যার সমানও হতে পারে।

প্রধান উপাদান নির্ণয় করার সময় এমনভাবে নির্ণয় করতে হবে যেন প্রথম প্রধান উপাদান উপাত্তের ভেদের বৃহত্তম অংশকেই ব্যাখ্যা করতে পারে। এই প্রথম প্রধান উপাদান হলো মূল চলকসমূহের ভারারোপিত রৈখিক সমাবেশ (weighted linear combination)। দ্বিতীয় প্রধান উপাদান হলো মূল চলক-সমূহের অন্য একটি ভারারোপিত রৈখিক সমাবেশ যা প্রথম রৈখিক সমাবেশের অনপেক্ষ এবং এটি মূল চলকের দ্বারা যে ভেদ ব্যাখ্যা করা যায়নি তার বৃহত্তর অংশকে ব্যাখ্যা করতে পারে। এভাবে m-তম ($m \leqslant p$, p হলো মূল চলকের সংখ্যা) প্রধান উপাদান নির্ণয় করা যায় যা তার পূর্ববর্তী রৈখিক সমাবেশসমূহের অনপেক্ষ এবং যা পূর্ববর্তী প্রধান উপাদানসমূহ দ্বারা যে ভেদ ব্যাখ্যা করা হয়নি তার বৃহত্তর অংশকে ব্যাখ্যা করতে পারে। এখানে m এর মান যত ছোট হয় ততোই ভাল।

ধরা যাক $X_1, X_2, \ldots, X_p$ হলো দৈব ভেক্টর X এর P উপাদান। অনুমান করা যাক যে $E(X)=0$ এবং $V(X)=\Sigma=(\sigma_{ij})$ এখানে Σ হলো বাস্তব ধনাত্মক সেমিডেফিনিট ম্যাট্রিক্স (real positive semidefinite matrix)। মনে করি $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p$ হলো Σ এর আইগেন মান এবং $\gamma_1, \gamma_2, \ldots, \gamma_p$ হলো যথাক্রমে আইগেন মানসমূহের আইগেন ভেক্টর। এই আইগেন ভেক্টরসমূহ দ্বারা গঠিত একটি ম্যাট্রিক্স $\Gamma=[\gamma_1\ \gamma_2\ \ldots\ \gamma_p]$ বিবেচনা করা যাক। এখানে Γ এর অর্ডার হলো $p \times p$ এবং λ_i গুলো বিভিন্ন হলে Γ হবে সমকৌণিক (orthogonal) ম্যাট্রিক্স। Γ সমকৌণিক ম্যাট্রিক্স হওয়ার কারণে এটি Σ ম্যাট্রিক্সকে কৌণিক (diagonal) ম্যাট্রিক্স-এ পরিণত করে। এ কারণে লেখা যায়

$$\Lambda = \Gamma'\Sigma\Gamma \qquad (8.2.1)$$

অথবা $$\Sigma=\Gamma\Lambda\Gamma'$$

এখানে $\Lambda = \mathrm{diag}(\lambda_1\ \lambda_2\ \ldots\ \lambda_p)$।

এখন X ভেক্টরকে সমকৌণিক পরিবর্তনের মাধ্যমে লেখা যায়

$$Y=\Gamma'X \qquad (8.2.2)$$

এখানে Y হলো $Y_1, Y_2, \ldots, Y_p$ বিশিষ্ট একটি ভেক্টর এবং Y_i ($i=1, 2, \ldots, p$) হলো λ_i এর প্রাসঙ্গিক প্রধান উপাদান।

উপরিউক্ত পরিবর্তনের কারণে X এর সর্বমোট ভেদ Y এর দ্বারা প্রকাশিত হবে। কারণ

$$V(Y)=\Gamma' V(X)\Gamma=\Gamma'\Sigma\Gamma=\Lambda \qquad (৪.২.৩)$$

এই Y এর ভেদের সার্বিক পরিমাপ হলো $\mathrm{tr}\,\Lambda$ বা জেনারালাইজড ভেদাঙ্ক হলো $|\Lambda|$ । এখানে $\mathrm{tr}\,\Lambda=\sum_{i=1}^{p}\lambda_i$ । আবার, X এর ভেদের সার্বিক পরিমাপ হলো $\mathrm{tr}\,\Sigma$ বা জেনারালাইজড ভেদাঙ্ক $|\Sigma|$ যদি $\Sigma>0$ হয় । এখানে

$$\mathrm{tr}\,\Sigma=\mathrm{tr}\,\Gamma\Lambda\Gamma'=\mathrm{tr}\,\Gamma\Gamma'\Lambda=\mathrm{tr}\Lambda=\sum_{i=1}^{p}\lambda_i$$

এখান থেকে বুঝা যাচ্ছে যে X থেকে Y-তে পরিবর্তন করার ফলে X এর সর্বমোট ভেদের কোনো পরিবর্তন হচ্ছে না । এটি জেনারালাইজড ভেদাঙ্ক হতেও বুঝা যায়।

$$|\Sigma|=|\Gamma\Lambda\Gamma'|=|\Lambda|=\prod_{i=1}^{p}\lambda_i$$

$$|\Lambda|=\prod_{i=1}^{p}\lambda_i$$

নমুনা হতে এই বিশ্লেষণ করার সময় Σ এর পরিবর্তে নমুনা ভেদাঙ্ক সহ-ভেদাঙ্ক ম্যাট্রিক্স S ব্যবহার করতে হবে । সেক্ষেত্রে আইগেন মান হবে, ধরা যাক l_j এবং আইগেন ভেক্টর g_j, তার মানসমূহ হলো g_{ij} ।

সংশ্লেষাঙ্ক ম্যাট্রিক্স হতে প্রধান উপাদান নির্ধারণ (Extraction of Principal Component from Correlation Matrix) : চলকের ভেদাঙ্ক স্কেলের উপর নির্ভরশীল বলে ভেদাঙ্ক সহ-ভেদাঙ্ক ম্যাট্রিক্স ব্যবহার করে প্রধান উপাদান নির্ধারণ করা হলে তা স্কেল দ্বারা প্রভাবিত হবে । সে কারণে উপাদান নির্ধারণ করার সময় আদর্শায়িত চলক ব্যবহার করা শ্রেয় । আদর্শায়িত চলক ব্যবহার করা হলে Σ ম্যাট্রিক্স ρ, সংশ্লেষাঙ্ক ম্যাট্রিক্স-এ পরিণত হয় । ফলে উপাদান নির্ধারণের ক্ষেত্রে Σ এর পরিবর্তে ρ ব্যবহার করে একই নিয়মেই প্রধান উপাদান নির্ধারণ করা যায় ।

এই নির্ধারণের ক্ষেত্রে ধরা যাক $E(X)=E(Y)=0$ । তাহলে X ও Y এর সহ-ভেদাঙ্ক হবে

$$E(XY')=E(XX'\Gamma)=\Sigma\Gamma=\Gamma\Lambda\Gamma'\Gamma=\Gamma\Lambda$$

সুতরাং X_i ও Y_j এর সহ-ভেদাঙ্ক হবে $\gamma_{ij}\lambda_j$ । কিন্তু X_i ও Y_j এর ভেদাঙ্ক হলো যথাক্রমে σ_{ii} এবং λ_j । কাজেই X_i ও Y_j এর সংশ্লেষাঙ্ক হবে

$$\rho_{ij} = \gamma_{ij}\lambda_j/(\sigma_{ii}\lambda_j)^{1/2} = \gamma_{ij}(\lambda_j/\sigma_{ii})^{1/2}$$

কিন্তু Σ এর পরিবর্তে ρ ম্যাট্রিক্স ব্যবহার করা হলে $\sigma_{ii}=1$ । সুতরাং

$$\rho_{ij} = \gamma_{ij}\sqrt{\lambda_j}$$

উপরিউক্ত বিশ্লেষণ হতে বলা যায় Y_j দ্বারা X_i এর ভেদের ব্যাখ্যা করা অনুপাতের পরিমাণ হলো ρ_{ij}^2 । এখন Y_j গুলো অসংশ্লেষিত হওয়ার কারণে যে কোনো m প্রধান উপাদান দ্বারা ব্যাখ্যা করা ভেদের অনুপাত হবে

$$\rho_{im}^2 = \sum_{j\in m}\rho_{ij}^2 = \frac{1}{\sigma_{ii}}\sum_{j\in m}\lambda_j\gamma_{ij}^2$$

$$= \sum_{j\in m}\lambda_j\gamma_{ij}^2, \quad \text{যখন} \quad \Sigma = \rho$$

এই ρ_{im}^2 এর হর হলো X_i এর ভেদ এবং লব হলো m উপাদান দ্বারা ব্যাখ্যা করা X_i এর ভেদের পরিমাণ। যখন $m=p$, তখন লব হলো $\Gamma\Lambda\Gamma'$ ম্যাট্রিক্স এর (i, i)-তম মান, σ_{ii} । সেক্ষেত্রে $\rho_{im}^2=1$ । নমুনা হতে এই বিশ্লেষণ করার সময় ρ এর পরিবর্তে নমুনা সংশ্লেষাঙ্ক ম্যাট্রিক্স R ব্যবহার করতে হবে।

উপরিউক্ত পরিবর্তন হতে পাওয়া যায়

(i) $E(Y_i)=0$, (ii) $V(Y_i)=\lambda_i$, (iii) $Cov(Y_i, Y_j)=0$ $(i\neq j)$

(iv) $V(Y_1) \geqslant V(Y_2) \geqslant \cdots \geqslant V(Y_p) \geqslant 0$

(v) $$\sum_{i=1}^{p} V(Y_i) = \text{tr}\,\Lambda = \text{tr}\,\Sigma = \sum_{i=1}^{p}\lambda_i$$

(vi) $$\prod_{i=1}^{p} V(Y_i) = |\Sigma|$$ (vii) $\gamma_{(X_i, Y_j)} = \gamma_{ij}\sqrt{\lambda_j}/\sqrt{\sigma_{ii}}$

এখানে σ_{ii} হলো Σ ম্যাট্রিক্সের i-তম সারির i-তম স্তম্ভের মান। লক্ষ্য করা গিয়েছে যে, Y ও X এর জেনারালাইজড্ ভেদাঙ্ক সমান। এটি সত্য হয় যদি সব λ_i ভিন্ন

ভিন্ন হয় এবং $\lambda_i > 0$ হয়। কারণ λ_i গুলো ভিন্ন ভিন্ন না হলে Γ সমকৌণিক হবে না। ধরা যাক, λ_i গুলোর মধ্যে r-সংখ্যক ($r<p$) হলো ভিন্ন ভিন্ন। এক্ষেত্রে, Γ ম্যাট্রিক্সকে এককভাবে নির্ণয় করা যায় না এবং একে সমকৌণিক করার জন্য $A=(A_1, A_2, \cdots, A_r)$ দ্বারা প্রাক গুণন করতে হয়। এখানে $A_i (i=1, 2, \cdots, r)$ হলো t_i অর্ডারের সমকৌণিক ম্যাট্রিক্স। এরূপ ক্ষেত্রে Λ এর অর্ডার $p \times p$ হয় না। ফলে

$$|\Lambda| \neq \prod_{i=1}^{p} \lambda_i$$

অবশ্য এরূপ অবস্থায় প্রধান উপাদান বিশ্লেষণের উদ্দেশ্য ব্যাহত হয় না; কারণ Σ ম্যাট্রিক্স এর পদসংখ্যা (rank) $r < p$ হলে λ_i গুলোর শেষের $(p-r)$ মান শূন্য হবে এবং ঐক্ষেত্রে r-সংখ্যক প্রধান উপাদান পাওয়া যাবে এবং ঐ r প্রধান উপাদানই X এর সম্পূর্ণ ভেদ ব্যাখ্যা করতে পারে।

উদাহরণ ৪.২.১ঃ উপরে আলোচিত প্রধান উপাদান নির্ধারণ পদ্ধতি উদাহরণ ১.১ এর C_2 এর উপাত্তের ক্ষেত্রে প্রয়োগ করা যাক। উক্ত উপাত্ত হতে প্রাপ্ত সংশ্লেষাঙ্ক ম্যাট্রিক্স সারণি ৪.১-এ দেয়া হলো। এখানে সংশ্লেষাঙ্ক ম্যাট্রিক্স হতে প্রধান উপাদান নির্ধারণ করার কারণ হলো যে, চলকের ভেদাঙ্ক চলকের স্কেলের (scale) উপর নির্ভর করে। আলোচিত উদাহরণে চলকসমূহ একই স্কেলে পরিমাপ করা হয় নি বলে তাদের ভেদাঙ্কের পরিমাণ একইরূপ হবে না। ফলে বেশি ভেদাঙ্কবিশিষ্ট চলক বিশ্লেষণে বেশি গুরুত্ব পাবে। বাস্তবে তার গুরুত্ব বেশি নাও হতে পারে। বড় স্কেলে পরিমাপ করার কারণে ভেদাঙ্কের পরিমাণ বেশি হয় মাত্র।

সারণি ৪.১ঃ সংশ্লেষাঙ্ক ম্যাট্রিক্স।

	A	B	C	D	E	F
A	1.0000	−0.1000	−0.5735	0.1321	−0.2190	0.0512
B		1.0000	0.6064	−0.2570	0.6026	0.1622
C			1.0000	−0.4253	0.3958	−0.0076
D				1.0000	−0.1408	0.0218
E					1.0000	−0.2007
F						1.0000

আলোচিত উদাহরণের ক্ষেত্রে সংশ্লেষাঙ্ক ম্যাট্রিক্স হতে প্রাপ্ত আইগেন মান (Eigen value) এবং উপাদানসমূহ দ্বারা ব্যাখ্যা করা ভেদাঙ্কের পরিমাণ সারণি ৪.২-এ দেয়া হলো।

সারণি ৪.২ ঃ আইগেন মান এবং ব্যাখ্যা করা ভেদাঙ্কের পরিমাণ।

উপাদান	আইগেন মান λ_i	ভেদাঙ্কের % হিসাব	মোট ভেদাঙ্কের যোজিত % হিসাব
1	2.4581	40.9687	40.9687
2	1.1153	18.5885	59.5572
3	1.0201	17.0023	76.5595
4	0.8714	14.5238	91.0833
5	0.3840	6.4004	97.4837
6	0.1510	2.5163	100.0000

এই λ_i এর ভিত্তিতে প্রাপ্ত আইগেন ভেক্টর $\gamma_i (i = 1,2,\cdots, p)$ এর মান ব্যবহার করে মূল চলক X_i গুলোর রৈখিক সমাবেশই হলো প্রধান উপাদান। এখানে রৈখিক সমাবেশের জন্য ব্যবহৃত ভরগুলো সারণি ৪.৩-এ দেয়া হলো। দেখা যাচ্ছে যে, প্রথম উপাদান দ্বারা উপাত্তের মোট ভেদের 41% ভেদ প্রকাশিত হয়।

সারণি ৪.৩ ঃ প্রধান উপাদানসমূহের উপাদান ভর (component weights)।

চলক।	উপাদান-১	উপাদান-২	উপাদান-৩	উপাদান-৪	উপাদান-৫	উপাদান-৬
A	0.35378	0.27679	0.61017	−0.45004	−0.26906	0.38858
B	−0.49853	0.33096	0.40082	0.04935	−0.35041	−0.59671
C	−0.56374	0.01326	−0.25412	0.041555	−0.50004	0.60471
D	0.32794	−0.08583	0.26574	0.84167	−0.30754	0.10712
E	−0.44751	−0.17755	0.53866	0.16774	0.60077	0.29832
F	0.02640	0.88022	−0.20418	0.23824	0.31508	0.16368

এর পরের দুটি উপাদান প্রায় 36% ভেদ ব্যাখ্যা করে। প্রথম চারটি উপাদের মোট ভেদের 91% প্রকাশ করে থাকে। আরো লক্ষ্য করা যাচ্ছে যে, প্রথম প্রধান উপাদান চলক A, B, C, D এবং E দ্বারা বেশি প্রভাবিত। তবে চলক B, C এবং E এর প্রভাব ঋণাত্মক। আবার প্রধান উপাদানের উপর চলক A এবং D এর ধনাত্মক প্রভাব আছে। দ্বিতীয় উপাদান যা চলকের প্রায় 19% ভেদ ব্যাখ্যা করে থাকে, তা চলক F দ্বারা বেশি প্রভাবিত। উপাদান-৩ চলক A এবং E দ্বারা বেশি প্রভাবিত। শেষ তিনটি উপাদানও একটি চলক দ্বারা বেশি প্রভাবিত। এখানে সারণি ৪.৩-এ সব উপাদানের জন্য চলকসমূহের ভর দেয়া হয়েছে। এই ভর ব্যবহার করে চলকের রৈখিক সমাবেশ প্রধান উপাদানগুলো হলো :

$$y_1 = 0.35378A - 0.49853B - 0.56374C + 0.32794D - 0.44751E + 0.02640F$$

$$y_2 = 0.27679A + 0.33096B + 0.01326C - 0.08583D - 0.17755E + 0.88022F$$

$$y_3 = 0.61017A + 0.40082B - 0.25412C + 0.26574D + 0.53866E - 0.20418$$

$$y_4 = -0.45004A + 0.04935B + 0.041555C + 0.84167D + 0.16774E + 0.23824F$$

$$y_5 = -0.26906A - 0.35041B - 0.50004C - 0.30754D + 0.60077E + 0.31508F$$

$$y_6 = 0.38858A - 0.59671B + 0.60471C + 0.10712D + 0.29832E + 0.16368F$$

এখানে প্রতিটি উপাদানের সহগসমূহ আদর্শায়িত (standardized) করা আছে যার ফলে তাদের বর্গের যোগফল 1 ।

উপরিউক্ত প্রধান উপাদান নির্ধারণ পদ্ধতি ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্স ব্যবহার করেও করা যায়। সারণি ৪.৪-এ আলোচিত উপাত্তের ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্স উপস্থাপন করা হলো। এই ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্স হতে প্রাপ্ত আইগেন মান

সারণি 8.8 ঃ দিনে দুবার দোহন করা গরুর দুধ উৎপাদন ও অন্যান্য তথ্যভিত্তিক ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্স।

	A	B	C	D	E	F
A	21.526	−28.160	−171.500	82.710	−0.320	24.569
B		3680.260	2371.200	−2103.770	11.514	1017.414
C			4154.600	−3699.380	8.036	−50.430
D				18212.660	−5.986	304.792
E					0.099	−6.536
F						10694.597

এবং প্রধান উপাদানসমূহ দ্বারা প্রকাশিত ভেদের পরিমাণ সারণি 8.৫-এ দেয়া হলো। দেখা যাচ্ছে যে, প্রথম তিনটি উপাদান চলকসমূহের মোট ভেদের 96% প্রকাশ করতে পারে। এই বিশ্লেষণ হতে প্রাপ্ত উপাদানসমূহের ভর সারণি 8.৬-এ দেয়া হলো। এই ভরসমূহ হলো i-তম চলকের সাথে j-তম উপাদানের

সারণি 8.৫ ঃ ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্স ভিত্তিক আইগেন মান এবং ব্যাখ্যা করা ভেদের পরিমাণ।

উপাদান	আইগেন মান λ_i	ভেদাঙ্কের % হিসাব	মোট ভেদাঙ্কের যোজিত % হিসাব
1	19564.4	53.22	53.22
2	10849.1	29.51	82.73
3	4919.93	13.38	96.11
4	1418.32	3.86	99.97
5	11.965	0.03	100.00
6	0.051	0.00	100.00

সারণি ৪.৬ ঃ ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্স ভিত্তিক বিশ্লেষণ হতে প্রাপ্ত উপাদান-সমূহের ভর (component weight) ।

চলক	উপাদান-১	উপাদান-২	উপাদান-৩	উপাদান-৪	উপাদান-৫	উপাদান-৬
A	0.00652	0.00146	-0.02288	0.06608	0.99739	0.01652
B	-0.16319	0.14632	0.67133	0.70734	-0.03054	-0.00390
C	-0.25406	0.03433	0.66736	-0.69637	0.06304	0.00097
D	0.95317	0.01812	0.29505	-0.06367	0.00473	-0.00002
E	-0.00050	-0.00042	0.00247	0.00242	-0.01666	0.99986
F	0.01550	0.98847	-0.12792	-0.07945	0.00077	0.00095

সহভেদাঙ্ক । ভরগুলো হতে দেখা যাচ্ছে যে, প্রথম প্রধান উপাদান চলক D দ্বারা প্রভাবিত । দ্বিতীয় উপাদান প্রায় 30% ভেদ প্রকাশ করে এবং এটি চলক F দ্বারা খুব বেশি প্রভাবিত । তৃতীয় উপাদান মোট ভেদের 13% প্রকাশ করে এবং এটি চলক B এবং C দ্বারা বেশি প্রভাবিত ।

লক্ষ্য করা যাচ্ছে যে, উপাদান-১ এবং উপাদান-২ এমন দুটি চলক দ্বারা প্রভাবিত যেগুলোর ভেদাঙ্ক খুব বেশি । আবার তৃতীয় উপাদান চলক B এবং C দ্বারা বেশি প্রভাবিত । এ দুটি চলকের ভেদাঙ্কও মোটামুটি বেশি । এটি হতে বুঝা যাচ্ছে যে প্রধান উপাদান নির্ধারণ করার জন্য ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্স ব্যবহার করা হলে প্রথম কয়েকটি উপাদান বেশি ভেদাঙ্কবিশিষ্ট চলক দ্বারা বেশি প্রভাবিত হয়ে থাকে । এ কারণে ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্স ব্যবহার করার ক্ষেত্রে চলক-সমূহের ভেদাঙ্কের মধ্যে খুব বেশি পার্থক্য থাকা উচিত নয় । বাস্তবে চলকসমূহ বিভিন্ন স্কেলে পরিমাপ করার কারণে ঐগুলোর ভেদাঙ্কের মধ্যে বেশি পার্থক্য থাকা স্বাভাবিক । সে কারণেই প্রধান উপাদান বিশ্লেষণের ক্ষেত্রে সংশ্লেষাঙ্ক ম্যাট্রিক্স ব্যবহার করা উচিত ।

সারণি 8.৬-এ দেয়া ভর ব্যবহার করে প্রধান উপাদানসমূহকে লেখা যায় ঃ

$$y_1 = 0.00652A - 0.16319B - 0.2540C + 0.95317D - 0.00050E + 0.01550F$$

$$y_2 = 0.00146A + 0.14632B + 0.03433C + 0.01812D - 0.00042E + 0.98847F$$

$$y_3 = -0.02288A + 0.67133B + 0.66736C + 0.29505D + 0.00247E - 0.12792F$$

$$y_4 = 0.06608A + 0.70734B - 0.69637C - 0.06367D + 0.00242E - 0.07945F$$

$$y_5 = 0.99739A - 0.03054B + 0.06304C + 0.00473D - 0.01666E + 0.00077F$$

$$y_6 = 0.01652A - 0.00390B + 0.00097C - 0.00002D + 0.99986E + 0.00095F$$

এই বিশ্লেষণের ক্ষেত্রেও প্রতিটি উপাদানের সহগসমূহ আদর্শায়িত (standardized) করা, যার ফলে তাদের বর্গের যোগফল 1 । এখানে আদর্শায়িত ভর ব্যবহার করার কারণে $E(Y_i) = 0$ হলেও $V(Y_i) \neq l_i$ । কিন্তু সংশ্লেষাঙ্ক ম্যাট্রিক্স ব্যবহার করে বিশ্লেষণ করার ফলে লক্ষ্য করা যাচ্ছে যে $E(Y_i) = 0$ এবং $V(Y_i) = l_i$ । এছাড়া $\Sigma l_i = p$, এখানে p হলো চলকের সংখ্যা । সুতরাং আদর্শায়িত ভর আরোপ করে যে j-তম উপাদান পাওয়া গিয়েছে তা মোট ভেদাঙ্কের যে অংশ প্রকাশ করে তার মান হলো l_j/p । এখানে l_j হলো সংশ্লেষাঙ্ক ম্যাট্রিক্স হতে প্রাপ্ত j-তম আইগেন মান । আবার j-তম উপাদান-এর ক্ষেত্রে i-তম চলকের রৈখিক সমাবেশের ভর হলো $g_{ij}/\sqrt{l_j}$ । এটি আবার i-তম চলক ও j-তম উপাদানের সহ-ভেদাঙ্ক । বিশ্লেষণ যেহেতু সংশ্লেষাঙ্ক ম্যাট্রিক্স হতে করা, সে কারণে $g_{ij}\sqrt{l_j}$ বা $g_{ij}\sqrt{l_j}/\sqrt{S_{ii}}$ হলো i-তম চলক ও j-তম উপাদানের সংশ্লেষাঙ্ক বা j-তম উপাদানের ক্ষেত্রে i-তম চলকের ভর । এখানে S_{ii} হলো সংশ্লেষাঙ্ক ম্যাট্রিক্স-এর i-তম সারির মান এবং g_{ij} হলো j-তম আইগেন মানের ভিত্তিতে প্রাপ্ত আইগেন ভেক্টরের i-তম মান ।

এতক্ষণ প্রধান উপাদান নির্ধারণ পদ্ধতি আলোচনা করা হয়েছে । এখানে বিশ্লেষণ হতে প্রাপ্ত সকল উপাদান দেখানো হয়েছে । বাস্তবে চলকের সংখ্যার চেয়ে কমসংখ্যক উপাদান নির্বাচন করা যেতে পারে পরবর্তী কোনো বিশ্লেষণের জন্য । লক্ষ্য করা গিয়েছে যে, সংশ্লেষাঙ্ক ম্যাট্রিক্স হতে বিশ্লেষণ করার ফলে প্রথম চারটি উপাদান মোট ভেদাঙ্কের 91% প্রকাশ করতে পারে । কাজেই ছয়টি চলকের পরিবর্তে চারটি প্রধান উপাদান নির্ধারণ করা হলেই পরবর্তী বিশ্লেষণ করা যাবে । এখানে চারটি প্রধান উপাদান নির্ধারণ করার জন্য ব্যবহৃত ভরগুলো 8.৭ সারণিতে দেয়া হলো । এই ভরগুলো আদর্শায়িত নয় ।

সারণি ৪.৭ঃ সংশ্লেষাঙ্ক ম্যাট্রিক্সভিত্তিক বিশ্লেষণ হতে প্রাপ্ত উপাদানসমূহের ভর।

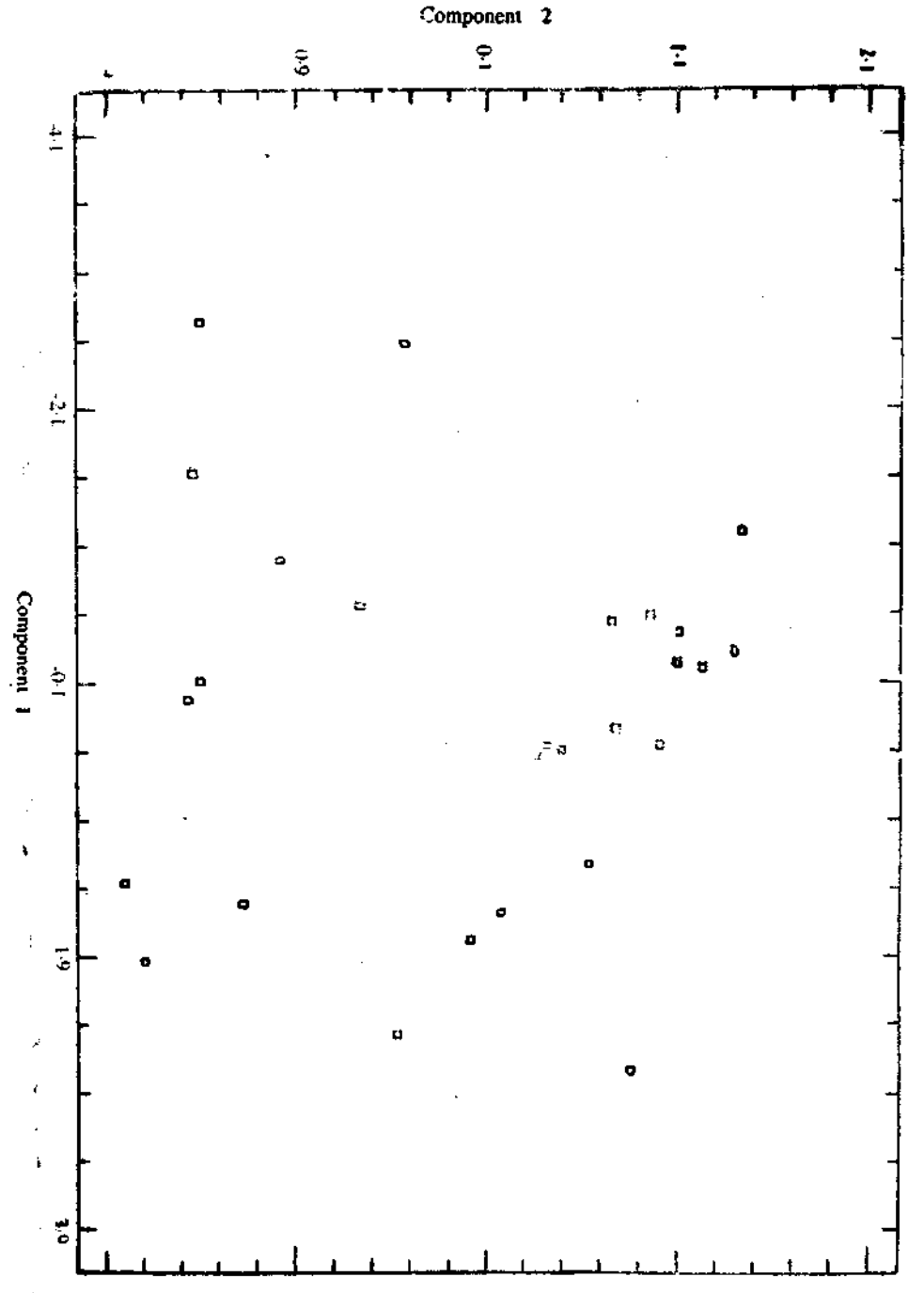
চলক	উপাদান-১	উপাদান-২	উপাদান-৩	উপাদান-৪
A	0.55468	0.29231	0.61628	−0.42011
B	−0.78161	0.34952	0.40484	0.04607
C	0.88385	0.01400	−0.25666	0.03879
D	0.51416	−0.09064	0.26840	0.78570
E	−0.70163	−0.18750	0.54406	0.15658
F	0.04139	0.92959	−0.20623	0.22240

৪.৩ চিত্রের মাধ্যমে প্রধান উপাদান উপস্থাপন (Graphical Representation of Principal Component)

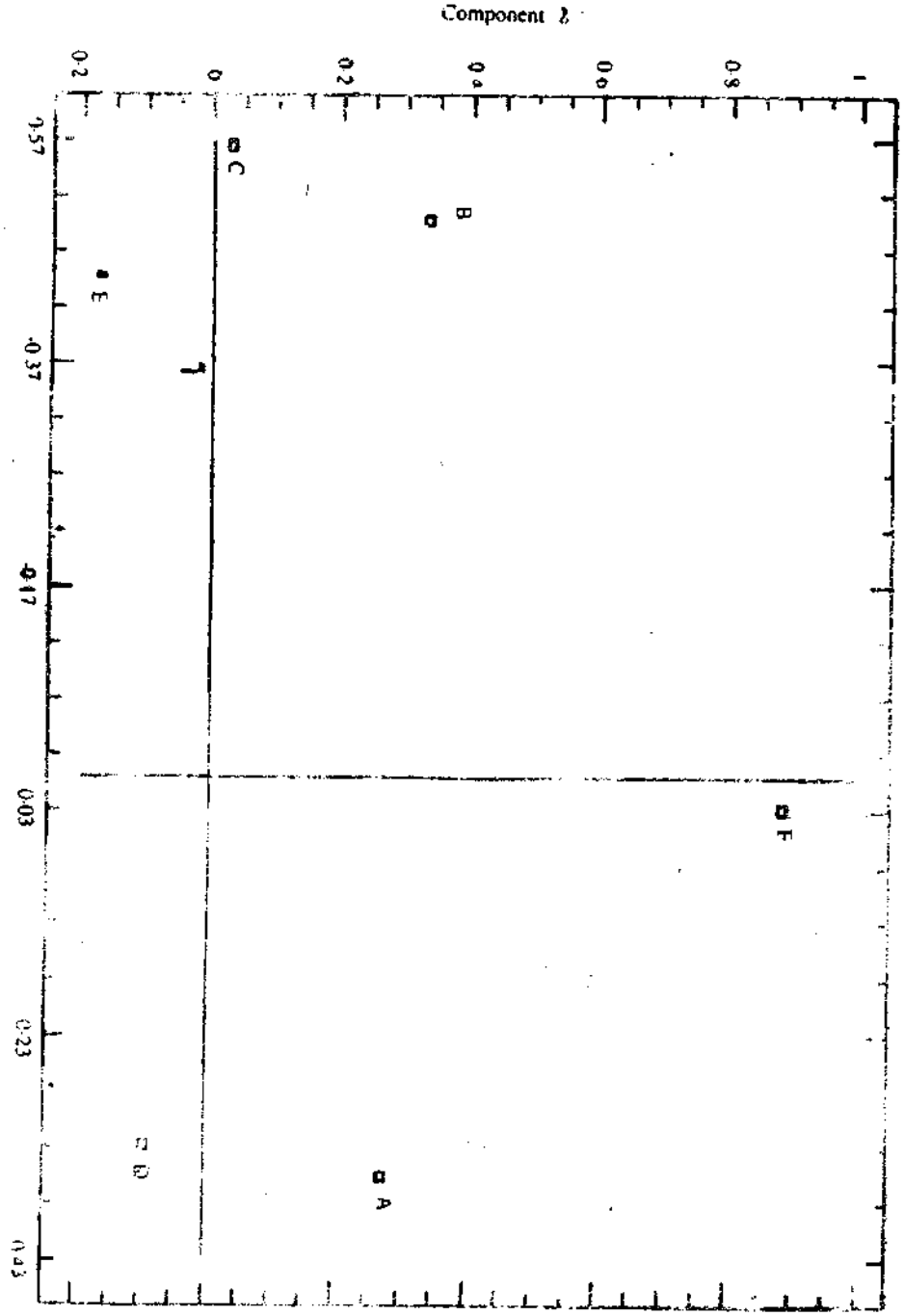
প্রধান উপাদান নির্ধারণের মাধ্যমে উপাত্ত সঙ্কোচিতকরণ পদ্ধতি চিত্রের মাধ্যমেও ব্যাখ্যা করা যায়। ধরা যাক প্রথম দুটি প্রধান উপাদান উপাত্তের মোট ভেদের একটি বৃহত্তর অংশ ব্যাখ্যা করে থাকে। তাহলে প্রধান উপাদানদ্বয়ের জন্য বিক্ষেপ বিন্দু (scatter plot) আঁকা হলে ঐ চিত্র উপাত্তের বিন্যাসের একটি ধারণা ব্যক্ত করে থাকে। এখানে ৪.১ চিত্রে প্রথম প্রধান উপাদানদ্বয়ের বিক্ষেপ বিন্দু দেখানো হলো। এই বিশ্লেষণ বিন্দুগুলো উদাহরণ ১.১-এর ক্ষেত্রে দিনে দুবার দোহন করা গরুর দুধ উৎপাদন সম্বন্ধীয় উপাত্তের ক্ষেত্রে সংশ্লেষাঙ্ক ম্যাট্রিক্স ভিত্তিক প্রধান উপাদান বিশ্লেষণ হতে প্রাপ্ত। লক্ষ্য করা যাচ্ছে যে, উপাদান-২ এর অক্ষ থেকে ভেদাঙ্ক পরিমাপ করা হলে তা উপাদান-১ এর অক্ষ থেকে পরিমাপ করা ভেদাঙ্ক অপেক্ষা বেশি হবে। অবশ্য উভয় অক্ষাভিমুখেই উভয় প্রধান উপাদান বেশ বিক্ষিপ্ত অবস্থায় আছে। অবশ্য এটি অপ্রত্যাশিত নয়। কারণ উভয় প্রধান উপাদানই সমকৌণিক (orthogonal)।

৪.২ চিত্রে প্রথম দুটি প্রধান উপাদানের ভরসমূহ উপস্থাপন করা হলো। এই চিত্র হতে লক্ষ্য করা যাচ্ছে যে, দুধ উৎপাদনের সাথে বেশি জড়িত চলকসমূহ, যেমন উৎপাদনের পরিমাণ (A), উৎপাদনকাল শেষ হওয়ার পর গরুর ওজন (C), উৎপাদনকাল শেষ হওয়ার পর গরুর শারীরিক অবস্থা (D) এবং মেসটিটিস (E) উপাদান-১ এর কাছাকাছিতে ঘনীভূত। এর মধ্যে C এবং E অক্ষের এক প্রান্তে এবং A ও D অক্ষের অপর প্রান্তে। অপরপক্ষে বাছুর প্রসবের সময় বয়স (F) অপর উপাদানের নিকটবর্তী।

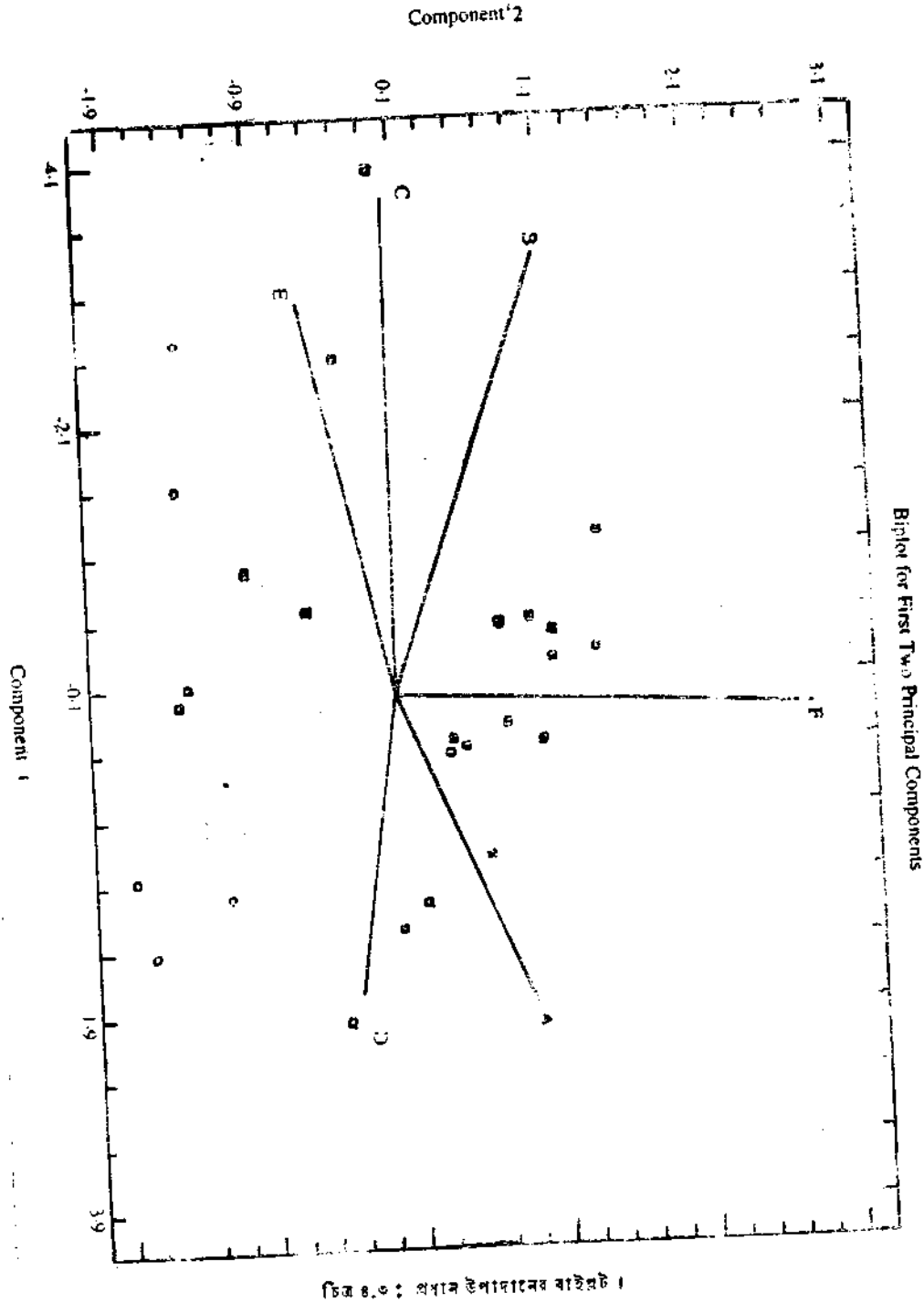
৪.৩ চিত্রে প্রথম প্রধান দুটি উপাদানের বাইপ্লট (biplot) উপস্থাপন করা হলো। এখানে ছয়টি রেখা (0, 0) বিন্দুতে মিলিত হয়েছে। এগুলো আদি চলক-সমূহ উপস্থাপন করছে। এখানে যে কোনো দুটি রেখার মধ্যে সৃষ্ঠ কোণ এদের সংশ্লেষাঙ্কের উল্টা সমানুপাতিক (inversely proportional)।



চিত্র ৪.১ : প্রধান উপাদানের বিক্ষেপ বিন্দু।



চিত্র ৪.২ : উপাদান ভরের চিত্র।



চিত্র ৪.৩ : প্রধান উপাদানের বাইপ্লট।

৪.৪ প্রধান উপাদানের ধর্মসমূহ (Properties of Principal Components)

ধরা যাক, X উপাত্ত ম্যাট্রিক্সের ভেদাঙ্ক ম্যাট্রিক্স হলো Σ, যার অর্ডার হলো $p \times p$ এবং Σ এর আইগেন মান হলো $\lambda_1, \lambda_2, \cdots, \lambda_p$ । ধরা যাক

$$\wedge = \text{diag}\,(\lambda_1 \cdots \lambda_p)$$

তাহলে Σ-কে লেখা যায়

$$\sum = \sum_{i=1}^{p} \lambda_i \gamma_i \gamma_i' = \Gamma \wedge \Gamma' \qquad (৪.৪.১)$$

এখানে γ_i গুলো হলো λ_i এর প্রাসঙ্গিক ভেক্টর এবং Γ হলো সমকৌণিক ম্যাট্রিক্স যার স্তম্ভগুলো হলো আদর্শায়িত আইগেন ভেক্টর । Σ ম্যাট্রিক্সকে উপরিউক্তভাবে বিযোজন করা হলে তাকে স্পেকট্রাল বিযোজন বলা হয় (spectral decomposition theorem) ।

আগেই উল্লেখ করা হয়েছে যে, i-তম প্রধান উপাদানের ভেদাঙ্ক হলো λ_i । এখন দেখানো যাবে যে, কোনো প্রধান উপাদানের ভেদাঙ্কই প্রথম উপাদানের ভেদাঙ্ক অপেক্ষা বড় নয় ।

ধরা যাক $a'X$ হলো একটি আদর্শায়িত প্রধান উপাদান । এটি X উপাত্ত ম্যাট্রিক্সের আদর্শায়িত রৈখিক সমাবেশ (Standardized Linear Combination, SLC) । এখানে

$$a'a = 1 \quad \text{এবং}$$

$$a = C_1\gamma_1 + C_2\gamma_2 + \cdots + C_p\gamma_p \text{ ।}$$

এখন $V(a'X) = a'\Sigma a$

$$= a' \left(\sum_{i=1}^{p} \lambda_i \gamma_i \gamma_i' \right) a$$

$$= \sum_{i=1}^{p} \lambda_i C_i^2 \qquad (৪.৪.২)$$

কিন্তু $a = C_1\gamma_1 + C_2\gamma_2 + \cdots + C_p\gamma_p$ এবং $a'a = 1$ হওয়ার কারণে $\Sigma C_i^2 = 1$ । আবার জানা আছে যে, $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$, অর্থাৎ λ_1 হলো সর্ববৃহৎ আইগেন মান এবং $\Sigma C_i^2 = 1$ হওয়ার শর্তে ৪.৪.২-এর সর্বোচ্চ মান হলো λ_1 । $V(a'X)$

সর্বোচ্চ হতে হলে $C_1 = 1$ এবং $C_2 = C_3 = \cdots = C_p = 0$ হতে হয়। তাহলে $a = \gamma_1$, অর্থাৎ $\acute{a}X = \gamma_1\acute{X}$। এখানে $\gamma_1\acute{X}$ হলো প্রথম প্রধান উপাদান।

উপপাদ্য 8.8.১ ঃ ধরা যাক $Z = \acute{a}X$ হলো একটি প্রধান উপাদান যা প্রথম k প্রধান উপাদানের সাথে অসংশ্লেষিত। এক্ষেত্রে Z এর ভেদাঙ্ক সর্বোচ্চ হবে যদি Z X-এর $(k+1)$-তম প্রধান উপাদান হয়।

প্রমাণ ঃ মনে করি $a = C_1\gamma_1 + C_2\gamma_2 + \cdots + C_p\gamma_p$; এখানে $\gamma_1, \gamma_2 \cdots, \gamma_p$ হলো $V(X) = \Sigma$ এর আইগেন ভেক্টর। শর্তানুসারে Z $\gamma_i\acute{X}(i = 1, 2, \ldots, k)$ এর সাথে অসংশ্লেষিত এবং সে কারণে $\acute{a}\gamma_i = 0$। সুতরাং $C_i = 0 (i = 1, 2, \cdots, k)$। আবার

$$V(Z) = \sum_{i=1}^{p} \lambda_i C_i^2$$

এবং $\Sigma C_i^2 = 1$। এখানে V(Z) এর সর্বোচ্চ মান হলো λ_{k+1}, কারণ $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \lambda_k \geqslant \lambda_{k+1} \geqslant \cdots \geqslant \lambda_p$ এবং $C_1 = C_2 = \cdots = C_k = 0$। সুতরাং $a = \acute{\gamma}_{k+1}$ এবং $Z = \acute{\gamma}_{k+1} X$।

প্রধান উপাদানের আরো কিছু ধর্ম

১। প্রধান উপাদান নির্ধারণ করার সময় r-সংখ্যক উপাদান নির্ধারণ করা যেতে পারে যদি Σ এর $(p-r)$ ছোট আইগেন মান সমান হয় $(0 < r < p-1)$। এই $(p-r)$ ছোট আইগেন মান সমান কিনা তা যাচাই করার জন্য Bartlett (1947) একটি যাচাই পদ্ধতি আলোচনা করেছেন। তাঁর নির্দেশিত যাচাই তথ্যজ-মান হলো

$$\chi^2 = M\left[-\log|S| + \sum_{j=1}^{r} \log l_j + (p-r)\log L\right] \qquad (8.8.৩)$$

এখানে l_j হলো নমুনা ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্স S এর আইগেন মান।

$$L = \frac{1}{p-r}\left(\text{tr } S - \sum_{j=1}^{r} l_j\right) \qquad (8.8.৪)$$

$$M = n - r - \frac{1}{6}\left[2(p-r) + 1 + \frac{2}{p-r}\right] \qquad (8.8.৫)$$

এই χ^2-এর বিন্যাস হলো $\frac{1}{2}(p-r-1)(p-r+2)$ স্বাধীনতার মাত্রাবিশিষ্ট প্রায় কাই-বর্গ বিন্যাস।

Lawley (1956) Bartlett's χ^2 এর একটি উন্নত আকার প্রস্তাব করেছেন। তাঁর মতে χ^2 এর সাথে

$$L^2 = \sum_{j=1}^{r} \frac{1}{(\lambda_j - L)^2} \qquad (8.8.6)$$

যোগ করতে হবে। সেক্ষেত্রে নতুন প্রস্তাবিত χ^2 এর স্বাধীনতার মাত্রা হবে

$$\tfrac{1}{2}(p-r+2)(p-r-1)$$

Anderson (1963) মধ্যমানের কিছু λ_j এর সমতা যাচাই পদ্ধতি আলোচনা করেছেন। ধরা যাক $\lambda_1, \lambda_2, \ldots, \lambda_q, \lambda_{q+1}, \ldots, \lambda_p$ হলো Σ ম্যাট্রিক্সের আইগেন মান। যাচাই করতে হবে

$$\lambda_{q+1} = \lambda_{q+2} = \ldots = \lambda_{q+k} = C \qquad (8.8.7)$$

এই নাস্তিকল্পনা যাচাই করার জন্য যাচাই তথ্যজমান হলো

$$X^2 = -(n-1)\sum_{j=q+1}^{q+k} \log \lambda_j + (n-1)k \log\left(\frac{1}{k}\sum_{j=q+1}^{q+k} \lambda_j\right) \qquad (8.8.8)$$

এই X^2-এর বিন্যাস হলো অভিসারী χ^2 বিন্যাস এবং এর স্বাধীনতার মাত্রা হলো $\frac{1}{2}[k(k+1)]-1$। এখানে লক্ষণীয় বিষয় হলো, যদি $q+k=p$ হয়, তাহলে X^2 যাচাই তথ্যজমান Bartlett প্রস্তাবিত Σ-এর শেষ m λ_j এর সমতা যাচাইয়ের যাচাই তথ্যজমানে পরিণত হয়। আবার, $q=0$ হলে X^2 এর আকার হয়

$$X^2 = -\left[(n-1) - \frac{1}{6}\left(2p+1+\frac{2}{p}\right)\right]\left[\log|S| + p\log\left(\frac{1}{p}\right)\sum_{j=1}^{p}\lambda_j\right] \qquad (8.8.9)$$

এই X^2 এর স্বাধীনতার মাত্রা হলো $(p-1)(p+2)/2$। এটি Bartlett প্রস্তাবিত $\lambda_1 = \lambda_2 = \ldots = \lambda_p$ যাচাইয়ের যাচাই তথ্যজমান। এটি Bartlett এর স্ফেরিসিটি (Sphericity) যাচাই নামে পরিচিত।

উপরে আলোচিত $(p-r)$ আইগেন মান সমান কিনা তা যাচাই করার জন্য সম্ভাব্যতা অনুপাত যাচাই (likelihood ratio test) পদ্ধতি প্রয়োগ করা যেতে পারে। ধরা যাক নাস্তিকল্পনা হলো

$$H_0 : \lambda_{r+1} = \lambda_{r+2} = \ldots = \lambda_p \qquad (8.8.10)$$

এই নাস্তিকল্পনা যাচাইয়ের মাধ্যমে r-সংখ্যক প্রধান উপাদান নির্ধারণ করা যেতে পারে। এখানে $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_r$ এবং উক্ত r-সংখ্যক আইগেন মান উপাত্তের অধিকাংশ ভেদ ব্যাখ্যা করার জন্য যথেষ্ট। বাকি $(p-r)$ আইগেন মানের প্রাসঙ্গিক প্রধান উপাদান উপাত্তের ভেদের সামান্য অংশই ব্যাখ্যা করতে পারে এবং নাস্তিকল্পনা সত্য হলে $(p-r)$ এর প্রতিটি উপাদানই সমান ভেদ ব্যাখ্যা করতে পারে। ফলে r-সংখ্যক প্রধান উপাদান নির্ধারণ করার পর নাস্তিকল্পনার অধীনে আর একটি প্রধান উপাদান নির্ধারণ করতে হলেই বাকি $(p-r)$ প্রধান উপাদানই নির্ধারণ করতে হবে। সুতরাং আলোচিত নাস্তিকল্পনা যাচাই করার সময় $r=0$ থেকে শুরু করে নাস্তিকল্পনা সত্য হওয়া পর্যন্ত যাচাই পদ্ধতি প্রয়োগ করে যেতে হয়। এক্ষেত্রে যাচাই তথ্যজমান হলো

$$-2 \log \lambda = np(a - 1 - \log g) \qquad (8.8.11)$$

এখানে a এবং g হলো $\hat{\Sigma}^{-1}S$ এর আইগেন মানসমূহের যথাক্রমে গাণিতিক ও জ্যামিতিক গড়, $\hat{\Sigma}$ হলো Σ এর সর্বাধিক সম্ভাবনা নিরূপক (Maximum likelihood estimator), S হলো নমুনা ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্স। কিন্তু $\hat{\Sigma}=S$ হওয়ার কারণে $\hat{\Sigma}$ ও S এর একই আইগেন মান এবং একই আইগেন ভেক্টর। তাছাড়া আরো জানা আছে যে, পরিমিত উপাত্তের ক্ষেত্রে $r>1$ হলে Σ এর আইগেন মানসমূহ ভিন্ন ভিন্ন হয় না বরং ঐগুলোর একটি সাধারণ মান হলো $\bar{\lambda}$। সেক্ষেত্রে

(i) $\bar{\lambda}$ এর সর্বাধিক সম্ভাবনা নিরূপক হলো $\bar{l}$, যেখানে $\bar{l}$ হলো নমুনা আইগেন মান-এর গাণিতিক গড়।

(ii) পুনর্বার প্রাপ্ত আইগেন মানের প্রাসঙ্গিক নমুনা আইগেন ভেক্টরসমূহ সর্বাধিক সম্ভাবনা নিরূপক। অবশ্য এগুলো একক সর্বাধিক সম্ভাবনা নিরূপক নয় [Anderson (1963), Kshirsagar (1972)]। কাজেই S এর আইগেন মান $(l_1, l_2, \ldots, l_p)$ হলে $\hat{\Sigma}$ এর আইগেন মান হবে $(l_1, l_2, \ldots, l_r, a_0, a_0, \ldots, a_0)$, যেখানে $a_0 = (l_{r+1} + \cdots + l_p)/(p-r)$ এবং এটি পুনর্বার প্রাপ্ত আইগেন মানের নমুনা নিরূপকের গাণিতিক গড়। এখানে অনুমান করা হচ্ছে যে পুনর্বার প্রাপ্ত আইগেন মানসমূহ শেষেই আসে, যদিও এ ধরনের অনুমানের তেমন কোনো প্রয়োজন নেই। উপরিউক্ত আলোচনার প্রেক্ষিতে বলা যায় যে, $\hat{\Sigma}^{-1}S$ এর আইগেন মানসমূহ হলো $(1, 1, \cdots, 1, l_{r+1}/a_0, \cdots, l_p/a_0)$। সুতরাং $a=1$ এবং $g=(g_0/a_0)^{(p-r)/p}$ বসিয়ে পাওয়া যায়

$$-2 \log \lambda = n(p-r) \log(a_0/g_0) \qquad (8.8.12)$$

এখানে $g_0 = (l_{r+1} \times \cdots \times l_p)^{1/(p-r)}$ = পুনর্বার প্রাপ্ত আইগেন মানসমূহের নমুনা নিরূপকসমূহের জ্যামিতিক গড়। এই শেষোক্ত $-2\log\lambda$ এর স্বাধীনতার মাত্রা হলো $(p-r-1)$। অবশ্য Σ যে সব সমকৌণিকতার শর্ত পূরণ করার কথা ঐগুলো H_0 দ্বারা ক্ষতিগ্রস্ত হয় বলে $-2\log\lambda$ এর স্বাধীনতার মাত্রা হলো $\frac{1}{2}(p-r+2)(p-r-1)$। এখানে $-2\log\lambda$ অভিসারী কাই-বর্গ বিন্যাস অনুসরণ করে। এই কাই-বর্গ বিন্যাস অধিক χ^2 বিন্যাসের কাছাকাছি হয় যদি n এর পরিবর্তে $n' = n-(2p+11)/6$ ব্যবহার করা হয়। অর্থাৎ

$$\left(n-\frac{2p+11}{6}\right)(p-r)\log\left(\frac{a_0}{g_0}\right) \sim \chi^2_{(p-r+2)(p-r-1)/2} \qquad (8.8.১৩)$$

লক্ষণীয় যে (a_0/g_0) এর মান S বা $S_s = \frac{n}{n-1}$ এর মান দ্বারা প্রভাবিত হয় না।

উদাহরণ 8.8.২ ঃ উদাহরণ 8.২.১-এর ক্ষেত্রে ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্স এর শেষ দুটি আইগেন মান সমান কিনা যাচাই করে দেখা যেতে পারে।

এখানে $p=6,\ r=4,\ n=28$

$$M = n-r-\frac{1}{6}\left[2(p-r)+1+\frac{2}{p-r}\right]$$

$$= 28-4-\frac{1}{6}[2\times 2+1+1] = 23$$

$$L = \frac{1}{p-r}\left[\text{tr } S - \sum_{j=1}^{r} l_j\right]$$

$$= \tfrac{1}{2}[36763.742 - 36751.75] = 5.996$$

$$|S| = 8.95\times 10^{14},\ \log|S| = 34.428$$

$$\chi^2 = M\left[-\log|S| + \sum_{j=1}^{r}\log l_j + (p-r)\log L\right]$$

$$= 23[-34.428 + 34.932 + 3.582]$$

$$= 93.978$$

এই χ^2 এর স্বাধীনতার মাত্রা হলো 2। এখানে $p(\chi^2 \geq 93.978) < 0.001$ হওয়াতে নাস্তিকল্পনা $\lambda_5 = \lambda_6$ বাতিল।

উপরের নাস্তিকল্পনা $H_0 : \lambda_6 = T_6$ যাচাই সম্ভাব্যতা অনুপাত যাচাই (LRT) এর মাধ্যমেও করা যায়। এখানে

$$a_0 = \tfrac{1}{6}(11.965 + 0.051) = 6.008, \quad g_0 = (11.965 \times 0.051)^{1/2} = 0.78$$

$$\left(n - \frac{2p+11}{6} = 28 - \frac{12+11}{6} = 24.17\right.$$

$$\therefore \quad -2\log\lambda = \left(n - \frac{2p+11}{6}\right)(p-r)\log(a_0/g_0)$$

$$= 24.17 \times 2 \times \log\left(\frac{6.008}{0.78}\right) = 98.69$$

এই $-2\log\lambda$ এর বিন্যাস হলো $\frac{1}{2}(p-r+2)(p-r-1) = 82$ স্বাধীনতার মাত্রাবিশিষ্ট χ^2। প্রাপ্ত χ^2-এর মান সারণিকৃত χ^2-এর 95-তম শতাংশকের বাইরে পড়ে বলে নাস্তিকল্পনা বাতিল। অর্থাৎ $\lambda_6 \neq \lambda_6$।

সংশ্লেষাঙ্ক ম্যাট্রিক্স হতে λ_j এর সমতা যাচাই : এতক্ষণ ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্স হতে প্রাপ্ত λ_j সম্পর্কে নাস্তিকল্পনা

$$H_0 : \lambda_{r+1} = \lambda_{r+2} = \quad = \lambda_p \qquad (8.8.18)$$

যাচাই পদ্ধতি আলোচনা করা হয়েছে। প্রধান উপাদান নির্ধারণ করার জন্য সংশ্লেষাঙ্ক ম্যাট্রিক্স $\rho - 3$ ব্যবহার করা হয়ে থাকে। এই ρ ম্যাট্রিক্স এর সকল আইগেন সমান যাচাই করা এবং নাস্তিকল্পনা $H_0 : \rho = I$ যাচাই করা একই কথা। এক্ষেত্রে অবশ্য 8.3.18 এর ন্যায় নাস্তিকল্পনা যাচাই করা সম্ভব হয় না। এমনকি এ ধরনের নাস্তিকল্পনা n বড় হলেও যাচাই করা কষ্টকর। সে যাই হোক $H_0 : \rho = I$ যাচাই করার জন্য যাচাই তথ্যজমান হবে।

$$\chi^2 = -\left[n - 1 - \frac{1}{6}(2p+5)\right]\log|R| \qquad (8.8.19)$$

এখানে R হলো সংশ্লেষাঙ্ক ম্যাট্রিক্স ρ এর নিরূপক। যাচাই তথ্যজমান 8.8.19 $\frac{1}{2}p(p-1)$ স্বাধীনতার মাত্রাবিশিষ্ট অভিসারী কাইবর্গ বিন্যাস অনুসরণ করে।

যাচাই তথ্যজমান 8.8.19 উদাহরণ 1.1-এ C_2 এর উপাত্তের ক্ষেত্রে প্রয়োগ করা যেতে পারে। উক্ত উপাত্তের জন্য R সারণি 8.1-এ দেয়া আছে। এখানে $|R| = 0.141285$, $n = 28$, $p = 6$। সুতরাং $\chi^2 = 47.29$। এই χ^2 এর স্বাধীনতার মাত্রা হলো 15। কিন্তু $p(\chi^2 \geq 47.29) < 0.001$ হওয়াতে বলা যায় যে R হতে প্রাপ্ত আইগেন মানগুলো সমান নয় বা $R \neq I$।

নাস্তিকল্পনা $H_0 : P = I$ যাচাই করার জন্য যাচাই তথ্যজমান 8.8.১৫ Box (1949) এর প্রস্তাব অনুসারে উপস্থাপিত। এই নাস্তিকল্পনা ছাড়া 8.8.১৪-ও যাচাই করা যেতে পারে। আগেই উল্লেখ করা হয়েছে এরূপ নাস্তিকল্পনা যাচাই সহজ নয়। তবু Bartlett (1951) এর প্রস্তাব অনুসারে LRT প্রয়োগ করে একটি যাচাই তথ্যজমান ব্যবহার করা যায়। এখানে

$$-2\log\lambda = (n-1)(p-r)\log(a_0/g_0) \qquad (8.8.১৬)$$

এখানে a_0 এবং g_0 হলো R ম্যাট্রিক্স হতে প্রাপ্ত ক্ষুদ্রতম আইগেন মানসমূহের যথাক্রমে গাণিতিক ও জ্যামিতিক গড়। অবশ্য যাচাই তথ্যজমান 8.8.১৬ অভিসারীভাবেও χ^2 বিন্যাস অনুসরণ করে না। তবে প্রথম r প্রধান উপাদান উপাত্তের মোট ভেদের একটি বৃহত্তর অংশ ব্যাখ্যা করে থাকলে 8.8.১৬-কে $\frac{1}{2}(p-r+2)\times(p-r-1)$ স্বাধীনতার মাত্রাবিশিষ্ট χ^2 তথ্যজমান হিসেবে বিবেচনা করা যেতে পারে (Dagnelie (1975), Anderson (1963))।

যাচাই তথ্যজমান 8.8.১৬ সারণি 8.১-এ দেয়া সংশ্লেষাঙ্ক ম্যাট্রিক্স হতে প্রাপ্ত আইগেন মানসমূহের সমতা যাচাই করার জন্য প্রয়োগ করা যেতে পারে।

এখানে $H_0 : \lambda_5 = \lambda_6$

বিবেচনা করা যাক। সেক্ষেত্রে

$$a_0 = \tfrac{1}{2}(0.3840 + 0.1510) = 0.2675, \quad g_0 = (0.384 \times 0.151)^{1/2} = 0.2408$$

তাহলে $$-2\log\lambda = (28-1)(6-4)\log\left(\frac{0.2675}{0.2408}\right) = 5.68$$

এখানে $\chi^2 = -2\log\lambda$ এর স্বাধীনতার মাত্রা হলো 2। কিন্তু $p(\chi^2 \geqslant 5.68) > 0.05$ হওয়াতে $\lambda_5 = \lambda_6$ বিবেচনা করা যায়। কিন্তু $H_0 : \lambda_4 = \lambda_5 = \lambda_6$ বাতিল হয়ে যায়। কারণ $p(\chi^2 \geqslant 19.24) < 0.05$। এখানে χ^2 এর স্বাধীনতার মাত্রা হলো 10,

$$a_0 = \frac{1}{3}(0.8714 + 0.3840 + 0.1510) = 0.4688,$$

$$g_0 = (0.8714 \times 0.384 \times 0.151)^{1/3} = 0.36969, \quad r = 3।$$

(২) প্রথম r প্রধান উপাদান দ্বারা ব্যাখ্যা করা মোট ভেদের পরিমাণ হলো

$$(\lambda_1 + \lambda_2 + \ \dots\ + \lambda_r)/(\lambda_1 + \lambda_2 + \ \dots\ + \lambda_p)$$

এখানে মোট ভেদের পরিমাপ হলো $\mathrm{tr}\,\Sigma$। সারণি 8.২-এ দেয়া λ_j গুলোর মান থেকে দেখা যাচ্ছে যে প্রথম চারটি প্রধান উপাদান মোট ভেদের

$$\frac{2.4581+1.1153+1.0201+0.8714}{2.4581+1.1153+1.0201+0.8714+0.384+0.151}=91.08\%$$

প্রকাশ করে থাকে।

(৩) দৈব ভেক্টরের জন্য প্রাপ্ত প্রধান উপাদানসমূহ উপাত্তের স্কেল দ্বারা প্রভাবিত হয়। ১.১ উদাহরণের C_2 এর উপাত্তসমূহ বিভিন্ন স্কেলে পরিমাপ করা হয়েছে। যেমন, A এর স্কেল হলো kg, F এর স্কেল হলো দিন ইত্যাদি। উক্ত উপাত্তের ক্ষেত্রে S ম্যাট্রিক্স হতে প্রাপ্ত প্রধান উপাদানসমূহ এবং R ম্যাট্রিক্স হতে প্রাপ্ত প্রধান উপাদানসমূহ একই নয়। এখানে R ম্যাট্রিক্স স্কেল মুক্ত। এতে বুঝা যায় যে, প্রধান উপাদান উপাত্তের স্কেল দ্বারা প্রভাবিত।

(৪) যদি Rank $(\Sigma)=r<p$ হয়, তাহলে X এর মোট ভেদ প্রথম r প্রধান উপাদান দ্বারা প্রকাশিত হয়। কারণ Rank $(\Sigma)=r$ হলে Σ এর শেষ $(p-r)$ আইগেন মানসমূহ হলো শূন্য। কাজেই গুণাবলি-2 হতে বলা যায় যে প্রথম r প্রধান উপাদান দ্বারা X এর মোট ভেদ প্রকাশিত হয়ে থাকে।

৪.৫ আইগেন মানের গুণাবলি (Properties of Eigen Values)

(১) ধরা যাক X হলো পরিমিত উপাত্ত ম্যাট্রিক্স যার ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্স হলো Σ। মনে করি নমুনা আকার হলো n। তাহলে Σ এর নিরপেক্ষ নিরূপক হলো $S_s=n(n-1)^{-1}S$ এবং এর স্বাধীনতার মাত্রা হলো $(n-1)$। সুতরাং S_s ও S এর আইগেন মানসমূহ সমান।

(২) ধরা যাক Σ হলো ধনাত্মক ডেফিনিট ম্যাট্রিক্স এবং এর আইগেন মানসমূহ ভিন্ন ভিন্ন। মনে করি, $M\sim W_p(m,\Sigma)$ এবং $U=m^{-1}M$। আবার $\Sigma=\Gamma\Lambda\Gamma'$ এবং $U=GLG'$, এখানে G হলো নমুনা হতে প্রাপ্ত আইগেন ভেক্টর-এর ম্যাট্রিক্স এবং $L=\text{diag}(l_1, l_2, \ldots, l_p)$; $l_j(j=1, 2, \ldots, p)$ হলো নমুনা ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্সের আইগেন মান। তাহলে $n\to\infty$ হলে

(i) $l\sim N_p(\lambda, 2\Lambda^2/m)$, এখানে l হলো l_j এর ভেক্টর, যেখানে $V(l_j)=2\lambda_j^2/m$ এবং l_j হলো U এর আইগেন মান।

(ii) $g_j\sim N_p(\gamma_j, V_j/m)$, এখানে

$$V_j=\lambda_j\sum_{i\neq j}\frac{\lambda_i}{(\lambda_i-\lambda_j)^2}\gamma_i\gamma_i',$$

g_j হলো l_j এর প্রাসঙ্গিক আইগেন ভেক্টর এবং γ_j হলো λ_j এর প্রাসঙ্গিক গণসমষ্টি আইগেন ভেক্টর।

(iii) g_i এর r-তম মান এবং g_j এর s-তম মান এর সহভেদাঙ্ক হলো

$$\lambda_i\lambda_j\gamma_{ri}\gamma_{sj}/m(\lambda_i-\lambda_j)^2$$

(iv) l এর মানসমূহ অভিসারীভাবে G এর মানসমূহের অনপেক্ষ। আই নমুনা আকার n হতে প্রাপ্ত S_g এর আইগেন মান $l_1, l_2, \ldots, l_p$ হলে

$$l_i \sim N(\lambda_i, 2\lambda_i^2/n-1) \quad \text{এবং}$$

$$\log l_i \sim N(\log \lambda_i, 2/(n-1)$$

সুতরাং λ_i এর নিশ্চয়তা-পরিক্ষেপ হলো

$$\log \lambda_i = \log l_i \pm Z\sqrt{2/(n-1)}$$

এখানে Z হলো আদর্শায়িত পরিমিত বিন্যাসের বর্জনীয় মান।

৪.৬ প্রধান উপাদানের সংখ্যা সম্পর্কে সিদ্ধান্ত (Decision About Number of Principal Components)

প্রধান উপাদানের সংজ্ঞা হতেই বুঝা যায় যে এটি উপাত্ত সঙ্কোচনের একটি পদ্ধতি। এই পদ্ধতিতে প্রধান উপাদানের সংখ্যা এমনভাবে নির্ধারণ করতে হয় যেন নির্ধারিত উপাদানসমূহ উপাত্তের ভেদের একটি বৃহত্তর অংশকে ব্যাখ্যা করতে পারে। স্বাভাবিকভাবেই কিছু উপাদানকে বাদ দিয়ে দিতে হয়। এখন প্রশ্ন হলো কোন উপাদানসমূহকে বাদ দিতে হবে এবং কয়টি উপাদান রাখতে হবে পরবর্তী বিশ্লেষণের জন্য।

উপাদানের সংখ্যা সম্পর্কে সিদ্ধান্ত নেয়ার অনেক পদ্ধতিই আছে। এই পদ্ধতিসমূহের মধ্যে কোনটি যাচাই পদ্ধতির এবং কোনটি চিত্র মাধ্যমের উপর নির্ভরশীল। আবার এ সব পদ্ধতিগুলো বিশ্লেষণের মাধ্যমের উপরও নির্ভরশীল। প্রধান উপাদান বিশ্লেষণ উপাত্তের ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্স এবং সংশ্লেষাঙ্ক ম্যাট্রিক্স উভয় মাধ্যমে করা যায়। এখানে উভয় মাধ্যমের জন্য উপাদানের সংখ্যা সম্পর্কে সিদ্ধান্ত নেয়ার পদ্ধতি আলোচনা করা হবে।

ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্স-এর ক্ষেত্রে (In Case of Variance Covariance Matrix) : ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্স ব্যবহার করে প্রধান উপাদান নির্ধারণ করা হলে উপাদানের সংখ্যা নির্ধারণ করার জন্য λ_j $(j=1, 2, \ldots, p)$ সম্পর্কে নাস্তিকল্পনা যাচাই-এর উপর নির্ভর করা যায়। কোনো λ_j এর মান তাৎপর্যপূর্ণভাবে শূন্য হলে তার প্রাসঙ্গিক প্রধান উপাদানকে পরবর্তী বিশ্লেষণের জন্য বাদ দেয়া যেতে পারে। সাধারণত ছোট আইগেন মানসমূহ সমান হলে তাদের প্রাসঙ্গিক প্রধান উপাদান বাদ দেয়া হয়। কারণ এই ছোট আইগেন মানসমূহ মূল উপাত্তের সামান্য ভেদই ব্যাখ্যা করতে পারে। এখানে ৪.৪ অনুচ্ছেদে আইগেন মানসমূহের তাৎপর্য যাচাই পদ্ধতি আলোচনা করা হয়েছে।

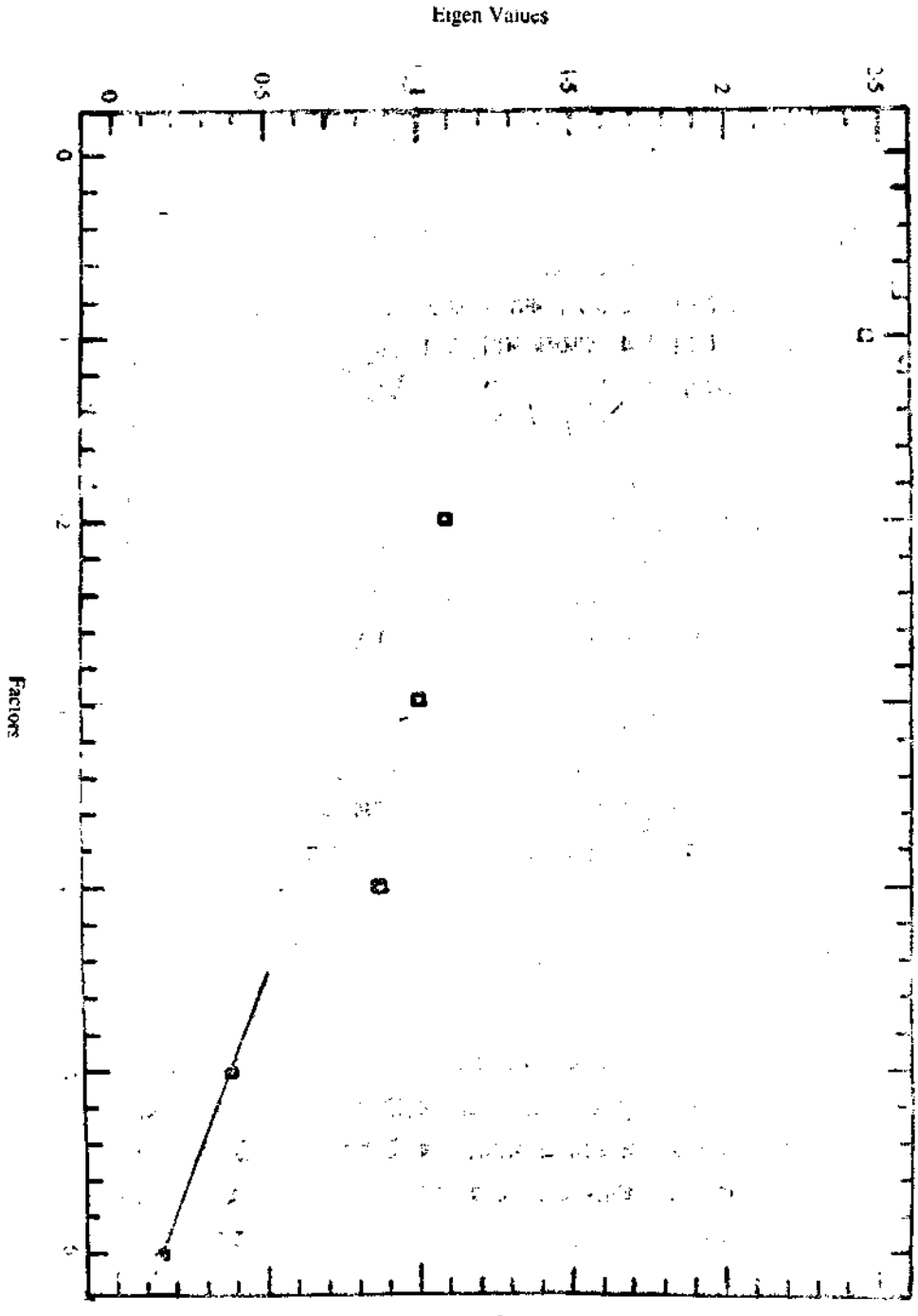
কিন্তু যাচাই পদ্ধতি নমুনা আকারের উপর নির্ভরশীল। নমুনা আকার মোটামুটি বড় হলেও অনেক উপাদানই তাৎপর্যপূর্ণ হয়। অবশ্য সব তাৎপর্যপূর্ণ উপাদানই যে মূল উপাত্তের ভেদের একটি বৃহত্তর অংশ ব্যাখ্যা করতে পারে তা নয়। কোনো কোনো তাৎপর্যপূর্ণ উপাদান ভেদের সামান্য অংশই ব্যাখ্যা করে থাকে। কাজেই সে সব তাৎপর্যপূর্ণ উপাদান বাদ দেয়াই বাস্তবসম্মত।

এই অবস্থায় অন্য কোনো পদ্ধতির কথা বিবেচনা করা যেতে পারে। এই পর্যায়ে মোট ভেদের বৃহত্তর অংশ যে উপাদানসমূহ দ্বারা ব্যাখ্যা করা যায় সে উপাদানগুলোই রাখা উচিত। ধরা যাক r উপাদান মোট ভেদের বৃহত্তর অংশ ব্যাখ্যা করে থাকে। এই ভেদের অংশ হলো

$$\sum_{j=1}^{r} l_j \Big/ \sum_{j=1}^{p} l_j .$$

এখানে l_j হলো নমুনা ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্স হতে প্রাপ্ত আইগেন মান। মোটামুটিভাবে যে সব আইগেন মান দ্বারা মোট ভেদের 90% বা তার বেশি প্রকাশিত হয় সে সব আইগেন মানের প্রাসঙ্গিক প্রধান উপাদানই পরবর্তী বিশ্লেষণের জন্য রাখা উচিত (Mardia et. al 1979)।

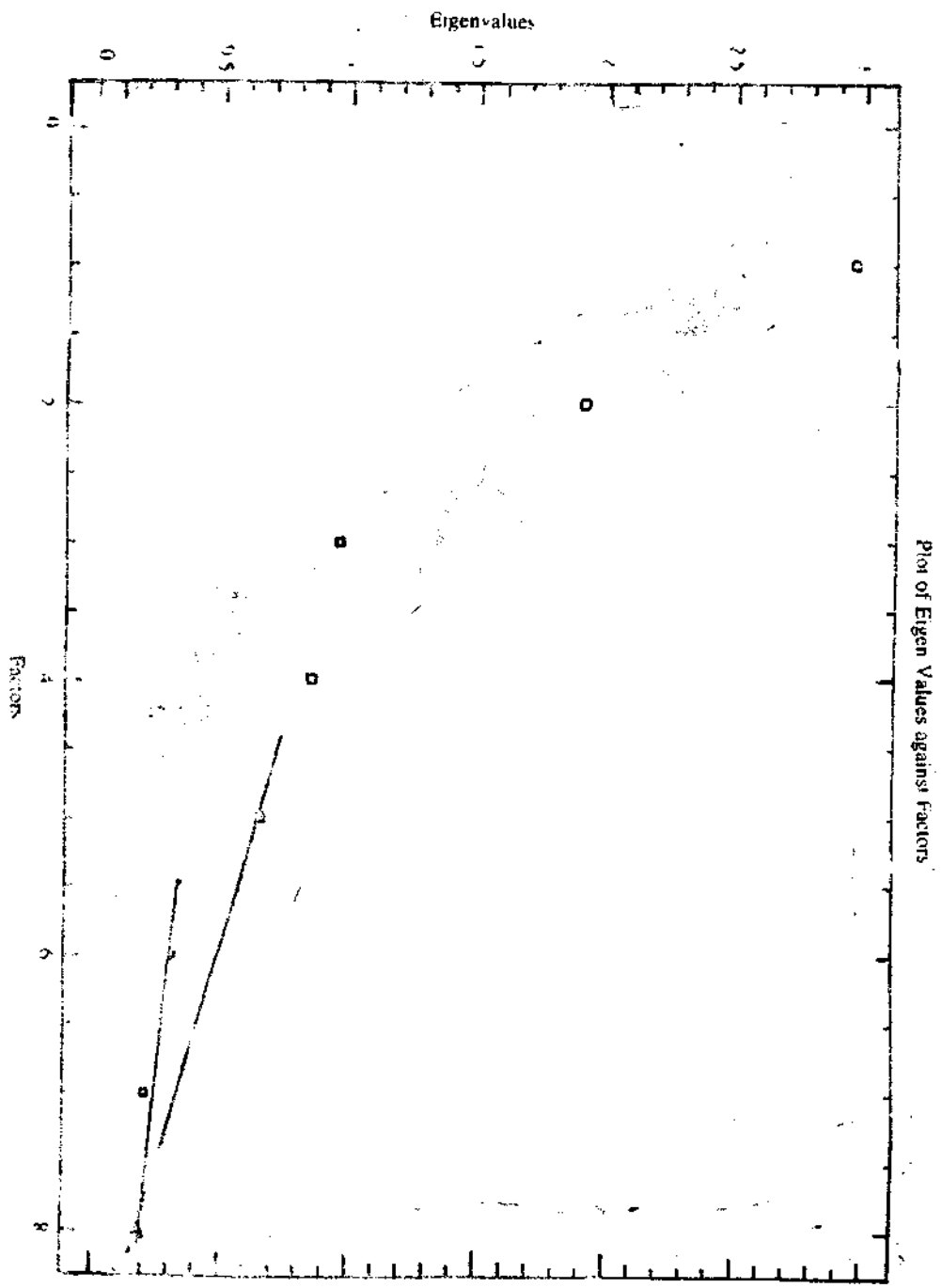
সংশ্লেষাঙ্ক ম্যাট্রিক্স এর ক্ষেত্রে (In Case of Correlation Matrix) : আগেই উল্লেখ করা হয়েছে যে, সংশ্লেষাঙ্ক ম্যাট্রিক্স হতে প্রধান উপাদান বিশ্লেষণ করা হলে উপাদানের সংখ্যা নির্ধারণ করার জন্য λ_j সম্পর্কে নাস্তিকল্পনা যাচাই করা অসুবিধাজনক। এ কারণে বিভিন্ন চিত্রের অধ্যায় উপাদানের সংখ্যা সম্পর্কে সিদ্ধান্ত নেয়ার উল্লেখ আছে। Cattell (1966) চিত্রের মাধ্যমে সিদ্ধান্ত নেয়ার একটি প্রস্তাব করেছেন। এই পদ্ধতিকে বলা হয় স্ক্রি-যাচাই (scree test)। এই পদ্ধতির জন্য X-অক্ষে উপাদান সংখ্যা এবং Y-অক্ষে উপাদানের প্রাসঙ্গিক আইগেন মান বসিয়ে একটি চিত্র আঁকতে হয়। ঐ চিত্রকে স্ক্রি-চিত্র (scree graph) বলা হয়। এখানে ১.১ উদাহরণের C_2-এর উপাত্তের ক্ষেত্রে সংশ্লেষাঙ্ক ম্যাট্রিক্স হতে প্রাপ্ত আইগেন মানের জন্য স্ক্রি-চিত্র 8.8-এ উপস্থাপন করা হলো। Cattell-এর প্রস্তাব অনুযায়ী স্ক্রি-চিত্রের ছোট আইগেন মানসমূহ যদি একটি সরলরেখার সৃষ্টি করে তাহলে ঐ সরলরেখার উপরে যে কয়টি বিন্দু থাকে ঐ বিন্দুর সংখ্যার সমান প্রধান উপাদানের সংখ্যা নির্ধারণ করতে হয়। পরে Cattell and Jaspers (1967) অনুরূপ প্রস্তাব করেছেন। তাঁদের প্রস্তাব অনুযায়ী চিত্র 8.8 লক্ষ্য করলে সিদ্ধান্ত নেয়া যায় যে ১.১ উদাহরণের C_2-এর উপাত্তের ক্ষেত্রে চারটি প্রধান উপাদান নির্ধারণ করা যায়।



চিত্র ৪.৪ : স্ক্রি চিত্র।

উপরিউক্ত স্ক্রি-যাচাই পদ্ধতি উপাদানের সংখ্যা সম্পর্কে সিদ্ধান্ত নেয়ার জন্য তুলনামূলকভাবে সহজ। কিন্তু এর কিছু অসুবিধাও আছে। প্রথমত কোন কোন বিন্দু দ্বারা সরলরেখা আঁকা হবে সে সম্পর্কে সিদ্ধান্ত নেয়া। দ্বিতীয়ত যে কোনো উপাত্তের ক্ষেত্রে একাধিক সরলরেখা আঁকা যেতে পারে। সে ক্ষেত্রে সরলরেখার শেষ কোথায় তার সিদ্ধান্ত নেয়া অসুবিধাজনক। যেমন, চিত্র ৪.৪-এ যে কেহ চতুর্থ ও ষষ্ঠ উপাদানের প্রাসঙ্গিক বিন্দু যোগ করেও একটি সরলরেখা পেতে পারে। আবার পঞ্চম ও ষষ্ঠ উপাদানের প্রাসঙ্গিক বিন্দু যোগ করেও একটি সরলরেখা পাওয়া যেতে পারে। এক্ষেত্রে সরলরেখার উপরের বিন্দুর সংখ্যা নির্ধারণ কষ্টকর। এটি চিত্র ৪.৫ থেকেও বুঝা যায়।

এতদসত্ত্বেও Horn (1965) উপাদানের সংখ্যা নির্ধারণ করার জন্য অন্য একটি প্রস্তাব করেছেন। তাঁর মতে বিশ্লেষণযোগ্য উপাত্তের জন্য একটি স্ক্রি-চিত্র আঁকতে হবে। পরে বহুচলক পরিমিত বিন্যাস হতে $(n \times p)$ অর্ডারের k গুচ্ছ নমুনা চয়ন করতে হবে। প্রতি নমুনার সংশ্লেষাঙ্ক ম্যাট্রিক্স আইডেনটিটি (identity) ম্যাট্রিক্স হতে হবে। এখন প্রতি নমুনার ক্ষেত্রে আইগেন মান নির্ণয় করে তাদের k গুচ্ছ নমুনার ভিত্তিতে আইগেন মানসমূহের গড় নির্ণয় করতে হবে এবং প্রতি উপাদানের প্রাসঙ্গিক গড় আইগেন মানকে উপাদানের বিপরীতে একই স্ক্রি-চিত্রে বসিয়ে চিত্রের বিন্দুগুলো যোগ করে দিতে হবে। এতে একটি সরলরেখার সৃষ্টি হবে। এই সরলরেখা বিশ্লেষণযোগ্য উপাত্তের জন্য আঁকা স্ক্রি-চিত্রের সাথে কোনো এক বিন্দুতে মিলিত হবে। ঐ মিলিত বিন্দুর উপরে প্রাপ্ত বিন্দুর সংখ্যার সমান প্রধান উপাদান নির্ধারণ করতে হবে। সাধারণত p/2 উপাদানের বিপরীতে যে বিন্দু থাকে ঐ বিন্দুতেই অঙ্কিত সরলরেখা মূল উপাত্তের স্ক্রি-চিত্রের সাথে মিলিত হয়। কারণ ঐ বিন্দুতে গড় আইগেন মানের মান হয় 1। এখানে আইগেন মান 1 বিবেচনা করার কারণ হলো কোনো কোনো গবেষক আইগেন মান 1 এর চেয়ে বড় হলে তার প্রাসঙ্গিক প্রধান উপাদান পরবর্তী বিশ্লেষণের জন্য নির্ধারণ করার প্রস্তাব করেছেন [Kaiser (1958)]।



চিত্র ৪.৫ : স্ক্রি চিত্র।

৪.৭ কিছু উপাদান বাদ দেয়ার প্রভাব (The Effect of Ignoring Some Components)

প্রধান উপাদান বিশ্লেষণ করার সময় কিছু উপাদানকে পরবর্তী বিশ্লেষণ হতে বাদ দেয়ার পদ্ধতি ৪.৬ অনুচ্ছেদে আলোচনা করা হয়েছে। এখন প্রশ্ন হলো কিছু উপাদান বাদ দেয়া হলে কি ক্ষতি হয়? এ প্রশ্নের উত্তর পেতে হলে লক্ষ্য করতে হবে কোন উপাদান কোন চলকের কত অংশ ব্যাখ্যা করতে পারে।

আগেই উল্লেখ করা হয়েছে যে Y_j দ্বারা X_i এর ভেদের ব্যাখ্যা করা অনুপাতের পরিমাণ হলো ρ_{ij}^2, এখানে

$$\rho_{ij}^2 = \lambda_j \gamma_{ij}^2 / \sigma_{ii}$$

এই তথ্যজমানের ভিত্তিতে নমুনার ক্ষেত্রে Y_j দ্বারা X_i এর ভেদের ব্যাখ্যা করা অংশের পরিমাণ হলো

$$r_{ij}^2 = l_j g_{ij}^2 / S_{ii}$$

এখানে γ_{ij} হলো j-তম আইগেন মানের প্রাসঙ্গিক আইগেন ভেক্টরের i-তম মান। অন্যভাবে বলা যায় γ_{ij} হলো j-তম উপাদানের ক্ষেত্রে রৈখিক সমাবেশের জন্য i-তম চলকের ভর। কোনো চলকের জন্য ভর যতো বেশি হবে ρ_{ij}^2 এর মান ততো বেশি হবে। অর্থাৎ i-তম চলকের ভেদের বেশি অংশ j-তম প্রধান উপাদান দ্বারা ব্যাখ্যা করা যাবে। ফলে j-তম উপাদান বাদ দেয়া হলে j-তম আইগেন মানের জন্য প্রাপ্ত আইগেন ভেক্টরের γ_{ij} মান বড় হলে i-তম চলকের অধিক তথ্য বাদ পড়ে যায়। বিষয়টি ১.১ উদাহরণের ক্ষেত্রে C_2 এর উপাত্তের জন্য সংশ্লেষাঙ্ক ম্যাট্রিক্স হতে বিশ্লেষণ করা প্রধান উপাদানের তথ্য হতে লক্ষ্য করা যাক।

আলোচিত উদাহরণের ক্ষেত্রে সকল উপাদানের জন্য r_{ij} ও r_{ij}^2 এর মান সারণি ৪.৮-এ দেয়া হলো। লক্ষ্য করলে দেখা যাবে যে, ভগ্নাংশের আসন্ন মান বাদ দিলে, r_{ij}^2 এর মানসমূহের সারির যোগফল 1। এই r_{ij} এর হতে বলা যায় প্রথম উপাদান চলক A এর 30.77%, B এর 61.09%, C এর 78.11%, D এর 26.44%, E এর 49.22% এবং F এর 0.17% ভেদ ব্যাখ্যা করতে পারে। প্রথম চারটি উপাদান চলক A এর 94.94%, B এর 89.91%, C এর 84.87%, D এর 95.83%, E এর 84.78% এবং F এর 95.79% ভেদ ব্যাখ্যা করতে পারে। এক্ষেত্রে শেষ দুটি উপাদান বাদ দেয়া হলে চলক C এবং E এর অধিক তথ্য নষ্ট হয়ে যাবে। অপরপক্ষে A, B, D এবং F এর ক্ষেত্রে তথ্য নষ্ট হবে কম।

সারণি ৪.৮ ঃ C_2 উপাক্তের j-তম উপাদানের সাথে i-তম চলকের সংশ্লেষাঙ্ক।

চলক	γ_{ij} এর মান উপাদান					
	1	2	3	4	5	6
1	0.5547	0.2923	0.6163	−0.4201	−0.1667	0.1510
2	−0.7816	0.3495	0.4048	0.0461	−0.2171	−0.2319
3	−0.8838	0.0140	−0.2567	0.0388	−0.3099	0.2350
4	0.5142	−0.0906	0.2684	0.7857	−0.1906	0.0416
5	−0.7016	−0.1875	0.5440	0.1566	0.3723	0.1159
6	−0.0414	0.9296	−0.2062	0.2224	0.1952	0.0636

γ_{ij}^2 এর মান

1	0.3077	0.0854	0.3798	0.1765	0.0278	0.0228
2	0.6109	0.1222	0.1639	0.0021	0.0471	0.0538
3	0.7811	0.0002	0.0659	0.0015	0.0960	0.0552
4	0.2644	0.0082	0.0720	0.6173	0.0363	0.0017
5	0.4922	0.0352	0.2959	0.0245	0.1386	0.0134
6	0.0017	0.8642	0.0425	0.0495	0.0381	0.0040

উপরিউক্ত আলোচনা হতে এটিই বুঝা যাচ্ছে যে, উপাদান বাদ দেয়ার সময় লক্ষ্য রাখতে হবে যে অধিকাংশ চলকের ভেদের বৃহত্তর অংশ যেন উপাদানসমূহ দ্বারা প্রকাশিত হয়। উপরের উদাহরণের ক্ষেত্রে প্রথম চারটি উপাদান চলক A,B,D এবং F এর মোট ভেদের 90% এর বেশি প্রকাশ করেছে। সেক্ষেত্রে চারটি প্রধান উপাদান নির্ধারণ মোটামুটি গ্রহণযোগ্য।

৪.৮ নির্ভরণে প্রধান উপাদান বিশ্লেষণ (Principal Component Analysis in Regression)

নির্ভরণ বিশ্লেষণে একটি সমস্যা হলো অনপেক্ষ (independent) চলকসমূহ সংশ্লেষিত (correlated) হতে পারে। সেক্ষেত্রে নির্ভরাঙ্কসমূহ যথাযথ (precise) হয় না। অনপেক্ষ চলকসমূহ সঠিকভাবে সংশ্লেষিত হলে $|X| = 0$ হয়, এখানে X হলো $n \times p$ অর্ডারের অনপেক্ষ চলকসমূহের উপাত্ত ম্যাট্রিক্স। ধরা যাক Y হলো নির্ভরশীল চলকের $n \times 1$ অর্ডারের ভেক্টর। y ও x_j ($j = 1, 2, \cdots, p$) চলকসমূহের মধ্যে একটি রৈখিক (linear) সম্পর্ক বিদ্যমান আছে অনুমান করা হলে ঐ সম্পর্ককে গাণিতিক প্রতিকৃতি (mathematical model)

$$Y = X\beta = U \qquad (8.৮.১)$$

দ্বারা প্রকাশ করা যায়। এখানে $\beta = [\beta_1\ \beta_2, \cdots, \beta_p]'$ হলো $(p \times 1)$ অর্ডারের পরামান ভেক্টর এবং U হলো $(n \times 1)$ অর্ডারের দৈব বিচ্যুতি (random error) ভেক্টর। এখন ন্যূনতম বর্গ পদ্ধতির (method of least squares) মাধ্যমে β ভেক্টরের একটি নিরূপক (estimate) পেতে হলে একটি অনুমান হলো $|X| \neq 0$ বা $\text{Rank}(X) = P$। সেক্ষেত্রে β এর নিরূপক হলো

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad (8.৮.২)$$

কিন্তু x_j চলকসমূহ সঠিকভাবে সংশ্লেষিত হলো $\text{Rank}(X) < P$ বা $|X| = 0$ হবে এবং $\hat{\beta}$ এর মান পাওয়া যাবে না। কারণ $(X'X)^{-1}$ পাওয়া যাবে না। আবার x_j চলকসমূহ সঠিকভাবে সংশ্লেষিত না হয়ে সাধারণভাবে সংশ্লেষিত হলে $|X| \neq 0$ হলেও তা শূন্য এর কাছাকাছি হবে [$|X| = 0$]। সেক্ষেত্রে $(X'X)^{-1}$ হয়ত পাওয়া যাবে এবং $\hat{\beta}$ এর মানও পাওয়া যাবে। কিন্তু $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$ বড় হবে। এরূপ অবস্থায় $\hat{\beta}$ এর যথাযথতা (precision) বাড়াতে হলে X ম্যাট্রিক্সকে সমকৌণিক (orthogonal) করতে হবে বা X ম্যাট্রিক্স-এর কিছু স্তম্ভ, অর্থাৎ x_j চলকসমূহের কিছু চলক বিশ্লেষণ হতে বাদ দিতে হবে। বর্তমান অনুচ্ছেদে প্রধান উপাদান বিশ্লেষণ পদ্ধতির মাধ্যমে এই চলক বাদ দেয়ার পদ্ধতি আলোচনা করা হবে।

ইতোমধ্যে ৪.৬ অনুচ্ছেদে উপাত্ত সঙ্কোচন পদ্ধতি হিসেবে প্রধান উপাদান বিশ্লেষণ পদ্ধতি আলোচনা করা হয়েছে। Massy (1965) প্রধান উপাদান নির্ধারণের

মাধ্যমে উপাত্ত সঙ্কোচনের উল্লেখ করেছেন। তবে ৪.৬ অনুচ্ছেদে আলোচিত উপাত্ত সঙ্কোচন পদ্ধতি হলো নির্ধারণ করা উপাদানগুলো যেন X ম্যাট্রিক্স-এর ভেদের একটি বৃহত্তর অংশ ব্যাখ্যা করতে পারে। কিন্তু নির্ভরণের ক্ষেত্রে প্রধান উপাদানগুলো এমনভাবে নির্ধারণ করতে হবে যেন Y ভেক্টরের সাথে সেগুলোর সংশ্লেষাঙ্ক বৃহত্তম হয়। Mittelhammer and Baritelle (1977) নির্ভরণ বিশ্লেষণের ক্ষেত্রে প্রধান উপাদান নির্ধারণ করার দুটি প্রক্রিয়ার উল্লেখ করেছেন। এখানে অন্যান্য পদ্ধতির সাথে ঐ দুটি পদ্ধতিও আলোচনা করা হবে।

বিশ্লেষণের সুবিধার্থে ধরা যাক $\bar{y} = \bar{x}_j = 0$ এবং $V(\bar{y}) = V(\bar{x}_j) = 1$, $j = 1, 2, \cdots, p$। মডেল ৪.৮.১ এর ক্ষেত্রে অনুমান করা যাক যে $U \sim N_n(0, \sigma^2 H)$, এখানে $H = I - n^{-1} ll'$। এখন X ভেক্টর-এর উপর সমকৌণিক পরিবর্তন $Z = \Gamma' X$ বিবেচনা করা যাক। এখানে Z হলো $Z_1, Z_2, \cdots, Z_p$ বিশিষ্ট একটি ভেক্টর, $\Gamma = (\gamma_1\ \gamma_2\ \cdots\ \gamma_p)$ ম্যাট্রিক্স, যেখানে γ_j হলো $n^{-1}(X'X)$ ম্যাট্রিক্স-এর আইগেন মান λ_j এর প্রাসঙ্গিক আইগেন ভেক্টর। ধরা যাক γ_{ij} হলো γ_j ভেক্টরের i-তম ($i = 1, 2, \cdots, p$) মান। এখানে $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ হলে Γ সমকৌণিক ম্যাট্রিক্স হবে। এখন মডেল ৪.৮.১-এ Z এর মান বসিয়ে পাওয়া যায়

$$Y = Z\Gamma'\beta + U$$
$$= ZB + U \qquad (৪.৮.১)$$

এখানে $B = \Gamma'\beta$, এখন ন্যূনতম বর্গ পদ্ধতির মাধ্যমে পাওয়া যায়

$$\hat{B} = (Z'Z)^{-1}Z'Y = n^{-1}\Lambda^{-1}Z'Y$$

আবার $$\hat{B} = (Z'Z)^{-1}Z'(ZB + U) = B + (Z'Z)^{-1}Z'U$$
$$= B, \quad \text{যখন } E(U) = 0$$

এখান থেকে পাওয়া যায়

$$V(\hat{B}) = E(\hat{B} - B)(\hat{B} - B)'$$
$$= E(Z'Z)^{-1}Z'UU'Z(Z'Z)^{-1}$$
$$= \Lambda^{-1}\sigma^2$$

$\therefore$ $\hat{B} \sim N_p(B, \Lambda^{-1}\sigma^2)$। এখান থেকে লেখা যায় $\hat{B}_i = n^{-1}\lambda_i^{-1}Z_i'Y$

এবং $V(\hat{B}_i) = \sigma^2/n\lambda_i$, $i = 1, 2, \cdots, p$।

এখানে নির্ভরণ বর্গসমষ্টি (SSR) এবং বিচ্যুতির বর্গসমষ্টি (SSE) হলো

$$SSR = \hat{B}'Z'Y \text{ এবং } SSE = Y'Y - \hat{B}'Z'Y = Y'Y - \sum_{i=1}^{p} n\lambda_i \hat{B}_i^2$$

সুতরাং σ^2 এর নিরূপক $MSE = (n-p-1)^{-1}\left[Y'Y - \hat{B}'Z'Y\right]$। কাজেই B_i সম্পর্কে নাস্তিকল্পনা $H_0 : B_i = 0$ যাচাই করার জন্য যাচাই তথ্যজমান হবে

$$t = \hat{B}_i / SE(\hat{B}_i)$$

উপরিউক্ত বিশ্লেষণ হতে আদি পরামান ভেক্টর β-এর নিরূপকও পাওয়া যেতে পারে। কারণ, $\Gamma'\beta = B$ হওয়ার কারণে $\hat{\beta} = \Gamma\hat{B}$ সুতরাং,

$$\hat{B} \sim N_p(\Gamma B, n^{-1}\Lambda^{-1}\Gamma'\Gamma\sigma^2) \text{। কাজেই}$$

$$E(\hat{\beta}_i) = \sum_{j=1}^{p} \gamma_{ij} B_j \text{ এবং } V(\hat{\beta}_i) = \sigma^2 \sum_{j=1}^{p} \gamma_{ij}^2/n\lambda_i$$

নমুনা হতে $\hat{\beta}_i$ ও $V(\hat{\beta}_i)$ এর নিরূপক পাওয়ার সূত্র হলো

$$\hat{\beta}_i = \sum_{j=1}^{p} g_{ij} \hat{B}_j$$

এবং

$$v(\hat{\beta}_i) = \hat{\sigma}^2 \sum_{j=1}^{p} g_{ij}^2/nl_i$$

এখানে l_i হলো $n^{-1}(X'X)$ ম্যাট্রিক্স-এর নমুনাভিত্তিক আইগেন মান এবং g_{ij} হলো l_i এর প্রাসঙ্গিক আইগেন ভেক্টরের i-তম মান।

উদাহরণ ৪.৪.৩ : শিশু মৃত্যু এবং জনউর্বরতা (fertility) সম্পর্কীয় Bhuyan (1995)-এর কাজ হতে কিছু তথ্য এখানে উপস্থাপন করা হলো। উক্ত তথ্যের ভিত্তিতে প্রধান উপাদান নির্ভরণ বিশ্লেষণ করা যেতে পারে।

সারণি ৪.৯ ঃ শিশু মৃত্যু এবং জনউর্বরতা সম্পর্কীয় তথ্য।

মোট জন্ম-গ্রহণ করা সন্তান	সামাজিক মর্যাদা	পিতার শিক্ষা	মাতার শিক্ষা	আকাংখিত সন্তানের সংখ্যা	মৃত সন্তানের সংখ্যা	পিতার পেশা	মাতার পেশা	বিবাহিত জীবন কাল
y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
3	1	0	0	3	0	4	0	8
3	1	0	0	3	0	4	0	9
2	1	6	0	3	0	1	0	7
3	1	0	0	3	0	4	0	11
3	1	0	0	3	0	4	0	7
3	1	0	0	3	0	4	0	8
2	1	0	0	2	0	4	0	5
2	1	0	0	2	0	4	0	7
2	1	0	0	2	0	2	0	8
2	1	0	0	3	0	4	0	5
4	1	0	0	3	1	4	0	10
2	1	0	0	2	0	4	0	7
2	1	0	0	3	0	4	0	6
2	1	0	0	3	0	1	0	6
2	1	0	0	2	0	1	0	7
2	1	0	0	2	0	1	0	7
2	1	0	0	2	0	1	0	7
3	1	0	0	2	1	1	0	9

3	1	0	0	3	0	1	0	7
3	1	4	0	3	0	1	1	8
7	1	2	0	10	2	0	1	17
8	1	2	0	10	1	1	1	18
6	1	2	0	5	2	1	1	15
3	1	5	0	2	0	1	1	4
11	1	2	0	10	4	1	1	21
1	1	2	0	2	0	1	1	3
4	1	2	0	3	0	0	1	11
8	1	1	0	10	1	1	1	19
7	1	1	0	8	2	1	1	19
3	2	0	0	3	0	1	0	10
4	2	0	0	3	1	4	0	11
2	2	4	0	3	0	4	0	5
3	2	0	0	3	0	4	0	12
5	2	0	0	3	2	4	0	12
6	2	0	0	4	2	4	0	15
2	2	0	0	3	0	4	0	5
3	2	0	0	3	0	2	0	12
3	2	0	0	3	0	1	0	7
3	2	0	0	3	0	1	0	8
3	2	0	0	3	0	1	0	15
2	2	0	0	8	2	1	0	10
2	3	10	5	2	0	2	0	7
5	3	0	0	4	1	1	0	12

4	3	0	0	4	0	1	0	11
3	3	0	0	3	1	1	0	7
5	3	0	0	4	2	1	0	15
5	3	0	0	4	0	1	0	15
3	3	0	0	3	0	1	0	8
5	3	0	0	3	2	1	0	10
2	3	5	0	3	0	1	0	6
3	3	2	0	3	0	3	0	8
1	3	8	3	2	0	1	1	5
3	3	0	0	7	1	4	2	8
12	3	0	0	10	4	2	2	22
6	3	0	0	8	1	4	2	20
8	3	0	0	10	1	2	2	20
3	3	0	0	3	0	1	1	14
3	3	0	0	3	0	2	1	14
2	3	0	0	2	0	2	2	11
4	3	3	2	5	0	1	0	8

সামাজিক মর্যাদা $(x_1) = 1$, নিম্ন ; 2, মধ্যম ; 3 উচ্চ। শিক্ষাগত যোগ্যতা (x_2, x_3) হলো স্কুলে পড়ার সময়কাল (বছরে)। পিতার পেশা : $x_6 = 0$, শ্রমিক ; $x_6 = 1$, কৃষক, $x_6 = 2$, চাকুরী, $x_6 = 3$, ব্যবসা, $x_6 = 4$, অন্যান্য। মাতার পেশা : $x_7 = 0$ গৃহিনী, $x_7 = 1$, কৃষি ; $x_7 = 2$, চাকুরী। বিবাহিত জীবনকাল (x_8) পরিমাপ করা হয়েছে সম্পূর্ণ বছরে।

উক্ত উপাত্তের ক্ষেত্রে x_j ($j = 1, 2, \ldots, 8$) গুলোর ম্যাট্রিক্স নিচে দেয়া হলো। ম্যাট্রিক্স হতে প্রাপ্ত আইগেন মানগুলো হলো $\lambda_1 = 2.96$, $\lambda_2 = 1.91$, $\lambda_3 = 0.95$, $\lambda_4 = 0.85$, $\lambda_5 = 0.64$, $\lambda_6 = 0.31$, $\lambda_7 = 0.20$ এবং $\lambda_8 = 0.18$ ।

সারণি ৪.১০ ঃ সংশ্লেষাঙ্ক ম্যাট্রিক্স ।

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	1.000	0.084	0.284	0.072	0.093	−0.114	0.204	0.196
x_2	0.084	1.000	0.720	−0.075	−0.114	−0.249	0.096	−0.219
x_3	0.284	0.720	1.000	−0.117	−0.127	−0.087	−0.028	−0.171
x_4	0.072	−0.075	−0.117	1.000	0.686	−0.170	0.585	0.786
x_5	0.093	−0.114	−0.127	0.686	1.000	−0.116	0.356	0.678
x_6	−0.114	−0.249	−0.087	−0.170	−0.116	1.000	−0.142	−0.168
x_7	0.204	0.096	−0.028	0.585	0.356	−0.142	1.000	0.521
x_8	0.196	−0.219	−0.171	0.786	0.678	−0.168	0.521	1.000

আইগেন মানের প্রাসঙ্গিক প্রধান উপাদানসমূহের ভরগুলো সারণি ৪.১১-এ দেয়া হলো।

সারণি ৪.১১ ঃ প্রধান উপাদানের ভরসমূহ।

উপাদান-১	উপাদান-২	উপাদান-৩	উপাদান-৪	উপাদান-৫	উপাদান-৬	উপাদান-৭	উপাদান-৮
0.1203	0.3248	0.7486	−0.4933	−0.0455	−0.1324	−0.2377	−0.0100
−0.1129	0.6238	−0.2421	0.2881	0.0392	−0.2219	−0.4674	0.4327
−0.1316	0.6194	0.1266	0.2705	−0.2436	0.3331	0.4916	−0.3126
0.5254	0.0361	−0.1023	0.1886	−0.0340	0.3197	−0.5479	−0.5225
0.4693	−0.0126	−0.0459	0.1588	−0.5335	−0.6487	0.2115	−0.0467
−0.1350	−0.2993	0.5879	0.7214	0.0168	−0.0176	−0.1209	0.1038
0.4020	0.1737	0.0560	0.1383	0.7954	−0.2314	0.3094	−0.0626
0.5285	−0.0196	0.0594	−0.0082	−0.1352	0.4953	0.1673	0.6519

উপরিউক্ত স্তরের ভিত্তিতে প্রথম প্রধান উপাদানের রৈখিক সমাবেশ হলো

$$z_1 = 0.1203x_1 - 0.1129x_2 - 0.1316x_3 + 0.5254x_4 + 0.4693x_5 - 0.135x_6 + 0.402x_7 + 0.5285x_8$$

অনুরূপভাবে অন্যান্য উপাদানগুলোকেও x_j সমূহের রৈখিক সমাবেশ হিসেবে প্রকাশ করা যায়। এই $Z_j (j = 1, 2, \ldots, 8)$ গুলো ব্যবহার করে মডেল ৪.৮.৩ মিল

সারণি ৪.১২ ঃ প্রধান উপাদানের ভিত্তিতে নির্ভরণ বিশ্লেষণের ফলাফল।

উপাদান-সমূহ	নিরূপক $\hat{B}_i$	পরিচিত বিচ্যুতি s. e($\hat{B}_i$)	তাৎপর্য মান	গড় বর্গ বিচ্যুতি, MSE	R^2
z_1	0.5239	0.0275	0.0000		
z_2	−0.0031	0.0342	0.9285		
z_3	−0.0626	0.0486	0.2034		
z_4	0.1202	0.0513	0.0229	0.1318	0.8838
z_5	−0.2558	0.0588	0.0001		
z_6	0.0939	0.0844	0.2708		
z_7	−0.0440	0.1059	0.6794		
z_8	0.2449	0.1121	0.0334		

করে প্রাপ্ত ফলাফল সারণি ৪.১২-এ দেয়া হলো। এই বিশ্লেষিত ফলাফল হতে পাওয়া যায়

$$\hat{y} = 0.5239\, z_1 - 0.0031\, z_2 - 0.0626\, z_3 + 0.1202\, z_4 - 0.2558\, z_5 + 0.0939\, z_6 - 0.0440\, z_7 + 0.2449\, z_8$$

আবার z_j গুলোর মান বসিয়ে পাওয়া যায়

$$\hat{y} = -0.0369\, x_1 + 0.0844\, x_2 - 0.0509\, x_3 + 0.2391\, x_4 + 0.3227\, x_5 + 0.0049\, x_6 - 0.0309\, x_7 + 0.5056\, x_8$$

এই শেষোক্ত রেখার নিরূপকগুলো $\hat{\beta}_j\ (j = 1, 2, \ldots, 8)$-এর পরিমিত বিচ্যুতি হলো

$$s.e\,(\hat{\beta}_1) = 0.0273, \quad s.e\,(\hat{\beta}_2) = 0.0339, \quad s.e\,(\hat{\beta}_3) = 0.0482$$

$$s.e\,(\hat{\beta}_4) = 0.0508, \quad s.e\,(\hat{\beta}_5) = 0.0583, \quad s.e\,(\hat{\beta}_6) = 0.0837$$
$$s.e\,(\hat{\beta}_7) = 0.1050, \quad s.e\,(\hat{\beta}_8) = 0.1112 \text{ ।}$$

Kendall (1957) উল্লেখ করেছেন যে, প্রধান উপাদান $z_j\,(j = 1, 2, \ldots, 8)$ ব্যবহার করে y এর নির্ভরণ রেখা মিল করা হলে j-তম উপাদান y-এর ভেদের $n\lambda_j\hat{B}_j^2$ পরিমাণ ব্যাখ্যা করতে পারে। নমুনা হতে এই ভেদের পরিমাণ নির্ণয় করার সূত্র হলো $nl_j\hat{B}_j^2$। সুতরাং j-তম উপাদান y-এর ভেদের $[nl_j\hat{B}_j^2/Y'Y]$ অংশ ব্যাখ্যা করতে পারে। এখানে $[nl_j\hat{B}_j^2/Y'Y]$ হলো j-তম উপাদান ও y এর সংশ্লেষাঙ্কের বর্গ $[r_{yz_j}^2]$।

উপরিউক্ত উদাহরণের ক্ষেত্রে $z_1, z_2, \ldots, z_8$ দ্বারা ব্যাখ্যা করা y এর ভেদের পরিমাণ হলো যথাক্রমে 48.695, 0.001, 0.223, 0.737, 2.536, 0.166, 0.023 এবং 0.640। শতকরা হিসাবে উপাদানগুলো y এর ভেদের 82.53, 0.002, 0.38, 1.25, 4.30, 0.28, 0.04 এবং 1.08 অংশ ব্যাখ্যা করতে পেরেছে।

এতক্ষণ প্রধান উপাদানের মাধ্যমে নির্ভরণ বিশ্লেষণ আলোচনা করা হয়েছে। এই বিশ্লেষণে সকল উপাদান ব্যবহার করা হয়েছে। বাস্তব প্রয়োগে কিছু উপাদান বাদ দিতে হয়। কারণ λ_j গুলো সব অসমান না হলে Γ ম্যাট্রিক্স সমকৌণিক হয় না। সেক্ষেত্রে কিছু λ_j বাদ দিয়ে বাকিগুলোর আইগেন ভেক্টর ব্যবহার করে Γ ম্যাট্রিক্স পেতে হয়। এখানে λ_j বাদ দেয়ার পদ্ধতি ৪.৬ অনুচ্ছেদে আলোচনা করা হয়েছে। Mittelhammer et al (1977) t-মাচাই এর মাধ্যমেও তাৎপর্যহীন B_j এর প্রাসঙ্গিক z_j বিশ্লেষণ হতে বাদ দেয়ার প্রস্তাব করেছেন। Bhuyan (1984) বাস্তব উপাত্তের ক্ষেত্রে λ_j ও t-মাচাইভিত্তিক z_j বাদ দিয়ে কিছু কিছু অনপেক্ষ চলকের অধিক যথাযথ নিরূপক লক্ষ্য করেছেন। এখানে λ_j ও t-মাচাইভিত্তিক z_j বাদ দিয়ে বিশ্লেষণ পদ্ধতি আলোচনা করা হবে।

ধরা যাক $\lambda_1 > \lambda_2 > \cdots > \lambda_r > \lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_p$ এবং λ_{r+1} থেকে λ_p পর্যন্ত আইগেন মানগুলো ছোট। এক্ষেত্রে r প্রধান উপাদান ব্যবহার করে মিল করা নির্ভরণ রেখা হলো

$$\hat{y}_1 = \sum_{j=1}^{r} \hat{B}_j z_j \qquad (৪.৮.৪)$$

এখানে $\hat{y}_1$ মিল করার ক্ষেত্রে বিচ্যুতির বর্গসমষ্টি হলে

$$SSE_1 = Y'Y = n\sum_{j=1}^{r} \lambda_j \hat{B}_j^2 \qquad (৪.৮.৫)$$

সমীকরণ ৪.৮.৪ এর ব্যবহৃত $\hat{B}_j$ এর মান পাওয়ার জন্য Γ ম্যাট্রিক্সকে বিভক্ত করে নিতে হয়। ধরা যাক Γ এর বিভক্তরূপ হলো $\Gamma=[\Gamma_r \Gamma_{p-r}]$। অনুরূপ z-কে লেখা যায় $z=[z_r z_{p-r}]$। তাহলে $\hat{B}_{sr}=(z_r' z_r)^{-1} z_r' Y$, এখানে $\hat{B}=[\hat{B}_r \hat{B}_{p-r}]$। বিশেষ করে

$$\hat{B}_i = n^{-1}\lambda_i^{-1} z_i' Y \; ; \; i=1, 2, \cdots, r \qquad (8.৮.৬)$$

ধরা যাক আদি চলক $x_1, x_2, \cdots, x_p$ এর নির্ভরাঙ্ক ভেক্টর হলো $\beta_r=[\beta_r \beta_{p-r}]'$। তাহলে $\hat{\beta}_r=\Gamma_r \hat{B}_r$। কিন্তু $\hat{\beta}_r$ নিখুঁতি নয়। এই নিরূপক ভেক্টরের i-তম মানের বোঁক হলো

$$\text{Bias}(\hat{\beta}_{ri}) = E(\hat{\beta}_{ri}) - \beta_i = -\sum_{j=r+1}^{p} \gamma_{ij} B_j$$

সুতরাং

$$\text{MSE}(\hat{\beta}_{ri}) = V(\hat{\beta}_{ri}) + [\text{Bias}(\hat{\beta}_{ri})]^2$$

$$= \sigma_r^2 \sum_{j=1}^{r} \gamma_{ij}^2 / r\lambda_i + \left[-\sum_{j=r+1}^{p} \gamma_{ij} B_j\right]^2$$

নমুনা থেকে প্রাপ্ত এই MSE এর মান হলো

$$\text{MSE}(\hat{\beta}_{ri}) = \hat{\sigma}_r^2 \sum_{j=1}^{r} g_{ij}^2 / n l_i + \left[-\sum_{j=r+1}^{p} g_{ij} \hat{B}_j\right]^2$$

সুতরাং $\hat{\beta}_{ri}$ এর তুলনায় $\hat{\beta}_r$ এর আপেক্ষিক দক্ষতা (Relative efficiency, RE) হলো

$$\text{RE}(\hat{\beta}_r) = \frac{\sigma_r^2 \sum_{j=1}^{r} \gamma_{ij}^2 / n\lambda_i + \left[-\sum_{j=r+1}^{p} \gamma_{ij} B_j\right]^2}{\sigma^2 \sum_{j=1}^{p} \gamma_{ij}^2 / n\lambda_i}$$

এখন উদাহরণ 8.8.3 এর ক্ষেত্রে উপাদান বাদ দেয়ার পদ্ধতি প্রয়োগ করে নির্ভরণ বিশ্লেষণ করা যাক। এখানে স্ক্রি-চিত্র (চিত্র 8.৫) অনুযায়ী প্রথম চারটি আইগেন মানের প্রাসঙ্গিক [$\lambda_1 = 2.96$, $\lambda_2 = 1.91$, $\lambda_3 = 0.95$ এবং $\lambda_4 = 0.85$] প্রধান উপাদান ব্যবহার করে বিশ্লেষণ করে যে ফলাফল পাওয়া গেছে তা সারণি 8.১৩-এ উপস্থাপন করা হলো। এখানে $\hat{\sigma}_r^2 = 0.1848$।

সারণি 8.১৩ ঃ স্ক্রি-চিত্রের মাধ্যমে প্রধান উপাদান বাদ দিয়ে নির্ভরণ বিশ্লেষণের ফলাফল।

উপাদান-সমূহ	নিরূপক $\hat{B}_{ri}$	পরিমিত বিচ্যুতি $s.e(\hat{B}_{rj})$	তাৎপর্য মান	নিরূপক $\hat{\beta}_{ri}$	MSE $\times(\hat{\beta}_{ri})$	R.E$(\hat{\beta}_r)$ %
z_1	0.5239	0.0325	0.0000	−0.0441	0.00080	107.34
z_2	−0.0031	0.0405	0.9396	−0.0113	0.01003	872.77
z_3	−0.0626	0 0575	0.2814	−0.0463	0.00162	69.73
z_4	0.1202	0.0607	0.0527	0.3042	0.00541	209 64
				0.2679	0.00418	122.98
				−0.0199	0.01018	145.32
				0.2232	0.06788	615.69
				0.2722	0.05938	480.21

এখানে লক্ষ্য করা যাচ্ছে যে λ_j ভিত্তিক প্রধান উপাদান বাদ দিয়ে নির্ভরণ বিশ্লেষণ করা হলে অধিকাংশ মূল চলকের প্রভাবই কম দক্ষভাবে নিরূপিত হয়। বর্তমান উদাহরণের ক্ষেত্রে কেবল চলক x_3 এর প্রভাব অধিক দক্ষভাবে নিরূপিত হয়েছে।

উপরিউক্ত উদাহরণের ক্ষেত্রে স্ক্রি-চিত্রের মাধ্যমে প্রধান উপাদান বাদ দিতে গিয়ে লক্ষ্য করা যাচ্ছে যে $\lambda_0 = 0.70$ এর ছোট λ_j গুলোর [Beale et al (1967), Jollife (1972)] প্রাসঙ্গিক প্রধান উপাদানসমূহ বাদ পড়েছে। Mittelhammer et al (1977) প্রধান উপাদান বাদ দেয়ার জন্য t-বাচাই পদ্ধতির প্রয়োগের উল্লেখ করেছেন। তাঁদের মতে y ও z_j ($j = 1, 2, \cdots, p$) এর নির্ভরণ বিশ্লেষণ করতে গিয়ে যেসব z_j এর প্রাসঙ্গিক B_j তাৎপর্যহীন হবে সেগুলো বিশ্লেষণ হতে বাদ দিতে হবে। আলোচিত উদাহরণের ক্ষেত্রে z_2, z_3,

z_6 এবং z_7 এর প্রভাব সারণি ৪.১২ তাৎপর্যহীন। তাই এই প্রধান উপাদান-গুলো বাদ দিয়ে পুনরায় y এর নির্ভরণ রেখা মিল করা হয়েছে। বিশ্লেষিত ফলাফল সারণি ৪.১৪-এ উপস্থাপন করা হলো। এই বিশ্লেষণ থেকে প্রাপ্ত বিচ্যুতির গড় বর্গসমষ্টি হলো $\hat{\sigma}^2_S = 0.1320$। এখানে S হলো t-যাচাই এর মাধ্যমে প্রাপ্ত তাৎপর্যপূর্ণ প্রধান উপাদানের সংখ্যা।

সারণি ৪.১৪ ঃ t-যাচাই-এর মাধ্যমে প্রধান উপাদান বাদ দিয়ে নির্ভরণ বিশ্লেষণের ফলাফল।

উপাদান-সমূহ	নিরূপক $\hat{B}_{Si}$	s.e$(\hat{B}_{Si})$	তাৎপর্য মান	নিরূপক $\hat{\beta}_{Si}$	MSE × $(\hat{\beta}_{Si})$	R.E$(\hat{\beta}_{Si})$ %
z_1	0.5239	0.0273	0.0000	0.0129	0.00267	358.25
z_4	0.1202	0.0508	0.0216	0.0715	0.00049	42.64
z_5	−0.2558	0.0583	0.0001	−0.0507	0.00057	24.53
z_8	0.2450	0.1112	0.0317	0.1786	0.00497	192.59
				0.3900	0.00634	186.53
				0.0371	0.00489	69.80
				0.0084	0.01057	95.87
				0.4702	0.01020	82.49

লক্ষ্য করা যাচ্ছে যে t-যাচাই এর মাধ্যমে প্রধান উপাদান বাদ দিয়ে নির্ভরণ বিশ্লেষণ করা হলে অধিকাংশ মূল চলকের প্রভাব অধিক দক্ষভাবে নিরূপণ করা যায়।

আলোচিত উদাহরণের ক্ষেত্রে λ_j ভিত্তিক এবং t-যাচাই ভিত্তিক উপাদান বাদ দিলে মাত্র একটি উপাদান z_1 পরবর্তী বিশ্লেষণের জন্য থাকে সারণি ৪.১৩। নিচে সারণি ৪.১৫-এ y ও z_1 এর নির্ভরণ বিশ্লেষিত ফলাফল উপস্থাপন করা হলো। এই বিশ্লেষণ থেকে প্রাপ্ত গড় বর্গ বিচ্যুতি হলো $\hat{\sigma}^2_k = 0.1915$, এখানে k হলো উভয় পদ্ধতির মাধ্যমে বাদ দেয়ার পর প্রধান উপাদানের সংখ্যা। এখানে প্রধান উপাদান z_1, y এর ভেদের 82.53% ব্যাখ্যা করতে পারে।

সারণি ৪.১৫ : λ_j ভিত্তিক ও t-যাচাইভিত্তিক প্রধান উপাদান বাদ দিয়ে নির্ভরণ বিশ্লেষণের ফলাফল।

উপাদান-সমূহ	নিরূপক $\hat{B}_{kl}$	s.e$(\hat{B}_{kl})$	তাৎপর্য মান	নিরূপক $\hat{\beta}_{kl}$	MSE × $(\hat{\beta}_{kj})$	RE$(\hat{\beta}_{kl})$ %
z_1	0.5239	0.0331	0.0000	0.0630		
				−0.0591		
				−0.0689		
				0.2753		
				0.2459		
				−0.0707		
				0.2106		
				0.2769		

একটি লক্ষণীয় বিষয় হলো t-যাচাই-এর মাধ্যমে প্রধান উপাদান বাদ দিলে কোনো কোনো চলকের অধিক দক্ষ নিরূপক পাওয়া যায় এবং উপাদানগুলো y চলকের 89.16% ভেদ ব্যাখ্যা করতে পেরেছে। অপরপক্ষে λ_j ভিত্তিক প্রধান উপাদান বাদ দেয়াতে বাকি উপাদানগুলো y এর ভেদের 84.16% ব্যাখ্যা করতে পারে। সবদিক বিবেচনা করে এক্ষেত্রে t-যাচাইভিত্তিক উপাদান বাদ দেয়া অধিক অর্থবহ।

৪.৯ চলক বাদ দেয়ার পদ্ধতি (Method of Discarding Variables)

প্রধান উপাদান বিশ্লেষণের ক্ষেত্রে উপাদান বাদ দেয়ার পদ্ধতি সম্পর্কে ৪.৬ অনুচ্ছেদে আলোচনা করা হয়েছে। উপাদান বাদ দেয়ার ক্ষেত্রে বিবেচনা করা হয় যে, যে উপাদানগুলো বিশ্লেষণের জন্য রাখা হবে সেগুলো যেন মূল চলক-সমূহের ভেদের 90% বা তার বেশি ব্যাখ্যা করতে পারে। কিন্তু যে উপাদান-গুলো রাখা হবে সেগুলোর প্রতিটিই মূল চলকসমূহের রৈখিক সমাবেশ। কাজেই কিছু উপাদান বাদ দিলেও পরবর্তী বিশ্লেষণ থেকে মূল চলক বাদ যায় না। চলক বাদ দেয়ার পদ্ধতি নির্ভরণ বিশ্লেষণে প্রয়োগ করা হয়। সে ক্ষেত্রে যে চলকসমূহ নির্ভরশীল চলকের 90% বা তার বেশি ভেদ ব্যাখ্যা করতে পারে সেগুলোই বিশ্লেষণের জন্য রাখা হয়। কোনো চলক গুচ্ছ 90% বা তার বেশি ভেদ ব্যাখ্যা

করতে পারে তা লক্ষ্য করার জন্য বহুল সংশ্লেষাঙ্কের বর্গ, R^2, নির্ণয় করা হয়। এ সম্পর্কে আরো জানার জন্য Beale et al (1967), Beale (1970) এবং Thompson (1978) পর্যালোচনা করা যেতে পারে। বর্তমান অনুচ্ছেদে প্রধান উপাদান বিশ্লেষণের মাধ্যমে চলক বাদ দেয়ার পদ্ধতি আলোচনা করা হবে।

ধরা যাক উপাত্ত ম্যাট্রিক্স X এর সংশ্লেষাঙ্ক ম্যাট্রিক্স ভিত্তিক প্রধান উপাদান বিশ্লেষণ করা হয়েছে এবং সংশ্লেষাঙ্ক ম্যাট্রিক্সের আইগেন মানসমূহ হলো $\lambda_1 > \lambda_2 > \cdots > \lambda_r > \; > \lambda_p$; এই বিশ্লেষণের মাধ্যমে $(p-r)$ চলক পরবর্তী কোনো বিশ্লেষণ হতে বাদ দিতে হবে। ধরা যাক r হলো জানা। এখন সবচেয়ে ছোট আইগেন মানের প্রাসঙ্গিক আইগেন ভেক্টর বিবেচনা করা যাক। এই আইগেন ভেক্টরের মানগুলো x_j $(j=1, 2, \cdots, p)$ চলকসমূহের রৈখিক সমাবেশের সহগ। এই সমাবেশের ক্ষেত্রে যে চলকের চিহ্নবর্জিত সহগ সবচেয়ে বড় তাকে প্রথমে বাদ দিতে হবে। তারপর পরবর্তী সবচেয়ে ছোট আইগেন মান বিবেচনা করতে হবে এবং তার প্রাসঙ্গিক আইগেন ভেক্টরের বড় মান যে চলকের সহগ হবে ঐ চলককে বাদ দিতে হবে। এই পদ্ধতি $(p-r)$ সংখ্যক চলক বাদ না দেয়া পর্যন্ত চলতে থাকবে।

উপরে আলোচিত r এর মান জানা না থাকলে কিভাবে r এর মান ঠিক করতে হবে। আগেই [(৪.৬) অনুচ্ছেদ] প্রধান উপাদান বাদ দেয়ার পদ্ধতি আলোচনা করা হয়েছে। সেখানেও উল্লেখ করা হয়েছে λ এর ছোট মানের প্রাসঙ্গিক প্রধান উপাদান বাদ দিতে হবে। λ এর ছোট মানের ব্যাপারে বিভিন্ন গবেষকের বিভিন্ন মত আছেন। Jolliffe (1972) $\lambda \leqslant 0.70$ হলে তার প্রাসঙ্গিক প্রধান উপাদান বাদ দেয়ার প্রস্তাব করেছেন। Jeffers (1967) বাস্তব উপাত্তের ক্ষেত্রে চলক বাদ দেয়ার জন্য $\lambda \leqslant 0.70$ হলে তাদের প্রাসঙ্গিক আইগেন ভেক্টরের বড় মানের প্রাসঙ্গিক চলক বাদ দিয়েছেন।

উদাহরণ ৪.৪.৩ এর ক্ষেত্রে লক্ষ্য করা যাচ্ছে যে $\lambda_5 = 0.64$, $\lambda_6 = 0.31$, $\lambda_7 = 0.20$ এবং $\lambda_8 = 0.18$। সুতরাং ঐ উপাত্তের ক্ষেত্রে চারটি চলক বাদ দেয়া যেতে পারে। উক্ত উপাত্তের ক্ষেত্রে পঞ্চম উপাদানের ভর 0.7954 (সারণি ৪.১১) সবচেয়ে বড় এবং এটি চলক x_7 এর সহগ। সুতরাং x_7 বাদ দেয়া যেতে পারে। উপাদান-৬ এর ক্ষেত্রে বড় সহগ 0.6487 হলো x_5 এর সহগ ; উপাদান-৭ এর ক্ষেত্রে 0.5479 হলো সবচেয়ে বড় মান এবং এটি x_4 এর সহগ ; উপাদান-৮ এর ক্ষেত্রে সবচেয়ে বড় মান হলো 0.6519 এবং এটি x_8 এর সহগ। সুতরাং আলোচিত উপাত্তের ক্ষেত্রে x_4, x_5, x_7 এবং x_8-কে বিশ্লেষণ হতে বাদ দিতে হবে। অবশ্য উক্ত পদ্ধতিতে চলক বাদ দিলেই যে পরবর্তী বিশ্লেষণ অর্থবহ তথ্য সরবরাহ

করবে তা নয়। কারণ আলোচিত উদাহরণের ক্ষেত্রে x_4, x_5, x_7 এবং x_8 বাদ দিয়ে x_1, x_2, x_3 এবং x_6 ব্যবহার করে নির্ভরণ বিশ্লেষণ করা হলে (সারণি ৪.১৬) শেষোক্ত চলকগুলো y চলকের ভেদের মাত্র 6.5% ব্যাখ্যা করতে পারে।

সারণি ৪.১৬ ঃ চলক বাদ দিয়ে নির্ভরণ বিশ্লেষণের ফলাফল।

চলক	নিরূপক $\hat{\beta}_i$	$s.e(\hat{\beta}_i)$	তাৎপর্য মান	R^2	তাৎপর্য মানসহ F	
					F	তাৎপর্য মান
x_1	0.1049	0.1396	0.45			
x_2	−0.0562	0.1993	0.78	0.065	0.95	0.44
x_3	−0.1564	0.2014	0.44			
x_6	−0.1704	0.1372	0.22			

দেখা যাচ্ছে যে উক্ত বিশ্লেষণ y চলকের ভেদ সম্পর্কে কোনো অর্থবহ তথ্যই সরবরাহ করতে পারেনি। ফলে উপরে আলোচিত চলক বাদ দেয়া পদ্ধতি সকল উপাত্তের জন্য যে সঠিক হবে এমন কথা নয়। এরূপ অবস্থায় Jolliffe (1972, 1973) প্রস্তাবিত প্রদ্ধতি অধিক গ্রহণযোগ্য হতে পারে। তাঁর মতে চলক এমনভাবে বাদ দিতে হবে বা r এর মান এমনভাবে নির্ধারণ করতে হবে যেন প্রধান উপাদানসমূহ উপাত্তের 80% ভেদ ব্যাখ্যা করতে পারে।

পঞ্চম অধ্যায়

উপাদান বিশ্লেষণ
(Factor Analysis)

৫.১ সূচনা (Introduction)

চতুর্থ অধ্যায়ে উপাত্ত সংকুচিত করার পদ্ধতি হিসেবে প্রধান উপাদান বিশ্লেষণ (Principal component analysis) পদ্ধতির উল্লেখ করা হয়েছে। যে কোনো গবেষণায় অনেক চলকের পরিমাপ করা হতে পারে। আবার একই চলকের বিভিন্ন একক ব্যবহার করেও পরিমাপ করা হতে পারে। চলকের সংখ্যা যত বেশি হবে ঐগুলোর দ্বি-চলক সংশ্লেষাঙ্কের সংখ্যাও তত বেশি হবে যা থেকে সহজে উপাত্ত সম্পর্কীয় কোনো সিদ্ধান্ত নেয়া অসুবিধাজনক। সামাজিক গবেষণায় একটি বিশেষ বৈশিষ্ট্য কতগুলো আন্তঃসম্পর্কীয় চলক দ্বারা প্রভাবিত। যেমন, জনউর্বরতা (Fertility) নির্ভর করে মা-বাবার শিক্ষার স্তর, তাদের পেশা, অর্থনৈতিক অবস্থা, মায়ের বিয়ের বয়স, আকাঙ্খিত সন্তানের সংখ্যা, মৃত সন্তানের সংখ্যা ইত্যাদির উপর এখানে জনউর্বরতার পার্থক্য পর্যালোচনা করার জন্য শুধু দম্পতিদের জীবিত জন্মগ্রহণ করা সন্তানের সংখ্যা বিশ্লেষণ করলে চলবে না। ঐ সাথে আন্তঃসম্পর্কীয় চলকগুলোতে কিরকম পরিবর্তন হচ্ছে তাও বিশ্লেষণ করতে হবে। এই বিশ্লেষণের মুখ্য উদ্দেশ্য হবে জনউর্বরতা এবং এর সাথে সম্পর্কিত উপাদানসমূহের সংশ্লেষণ কিরূপ তা জানা। এরূপ বিশ্লেষণ হতে গুরুত্বপূর্ণ চলকসমূহকেও চিহ্নিত করা যায় যা ভবিষ্যৎ বিশ্লেষণের জন্যও একটি সহজ দিক নির্দেশ করতে সহায়ক হবে। অর্থাৎ অনেক উপাদানের পরিবর্তে অল্পসংখ্যক উপাদান ব্যবহার করেও একগুচ্ছ চলকের সম্পর্ক বিশ্লেষণ প্রয়োজন হয়।

উপাদান বিশ্লেষণ হলো এমন একটি পরিসংখ্যানিক পদ্ধতি, যার মাধ্যমে অনেকগুলো আন্তঃসম্পর্কীয় চলকগুচ্ছের সম্পর্ক নির্দেশ করার জন্য তুলনামূলকভাবে অল্পসংখ্যক উপাদান ব্যবহার করা হয়। সুতরাং উপাদান বিশ্লেষণও প্রধান উপাদান বিশ্লেষণের ন্যায় একটি উপাত্ত সংকোচন পদ্ধতি। কিন্তু উপাদান বিশ্লেষণ ও প্রধান উপাদান বিশ্লেষণের মধ্যে পার্থক্য হলো যে, প্রধান উপাদানসমূহ এমনভাবে নির্ণয় করতে হয় যেন ঐগুলো আদি চলকসমূহের ভেদাঙ্কের বৃহত্তম অংশ ব্যাখ্যা করতে পারে। অপরপক্ষে উপাদান বিশ্লেষণের ক্ষেত্রে অল্পসংখ্যক উপাদান নির্ণয় করা হয় যা মূল চলকসমূহের সম্পর্ক পর্যালোচনা করতে সাহায্য করে। যেমন ধরা যাক, সন্তান উৎপাদন সংখ্যা পর্যালোচনা করার জন্য দম্পতির আর্থ-সামাজিক

চলকসমূহ পর্যালোচনা করা হলে লক্ষ্য করা যাবে যে মায়ের শিক্ষা, বাবার শিক্ষা এবং উভয়ের বিবাহের বয়স পরস্পর সম্পর্কিত এবং এগুলোর সমন্বয়ে একটি উপাদান হতে পারে। উভয়ের পেশা এবং তাদের আর্থিক অবস্থা পরস্পর সম্পর্কিত বলে উক্ত চলকগুলো আরেকটি উপাদান নির্দেশ করতে পারে। আবার আকাঙ্খিত সন্তানের সংখ্যা এবং মৃত সন্তানের সংখ্যা পরস্পর সম্পর্কিত বলে তারা তৃতীয় আরেকটি উপাদান নির্দেশ করতে পারে। কাজেই জনউর্বরতার পার্থক্য পর্যালোচনা করার জন্য সকল চলক ব্যবহার না করে শেষোক্ত তিনটি উপাদান দ্বারাই জনউর্বরতা কি ভাবে প্রভাবিত হচ্ছে তা পর্যালোচনা করা যায়।

বিষয়টি আরো ব্যাখ্যা করার জন্য Spearman (1904) এর একটি কাজ উদ্ধৃত করা যায়। ঐ বিশ্লেষণে ছেলে-মেয়েদের Classics (x_1), French (x_2) এবং English (x_3) পরীক্ষার ফলাফল পর্যালোচনা হয়েছে। উক্ত বিষয়গুলোর পরীক্ষার ফলাফল বিষয়ের উপর ছেলে-মেয়েদের দক্ষতা এবং তাদের সাধারণ দক্ষতা এর উপর নির্ভর করে। সুতরাং প্রতি ছেলে-মেয়ের ফলাফলকে নিম্নরূপভাবে প্রকাশ করা যায় :

$$x_1 = \lambda_1 f + u_1,\ x_2 = \lambda_2 f + u_2 \text{ এবং } x_3 = \lambda_3 f + u_3$$

এখানে f হলো ছেলে-মেয়েদের সাধারণ দক্ষতা যা সাধারণ উপাদান (Common factor) হিসেবে বিবেচিত এবং λ_1, λ_2 ও λ_3 হলো উপাদান ভর (Factor loadings), $u_i (i = 1, 2, 3)$ হলো দৈব বিচ্যুতি। এখানে x_i এবং f এর সম্পর্ক যত বেশি হবে u_i এর ভেদাঙ্ক তত কম হবে। এখানে u_i এর ভেদকে দুইভাগে বিভক্ত করা যেতে পারে। প্রথমত, একটি ছাত্রের সাধারণ দক্ষতা তার কোনো বিষয়ের দক্ষতার চেয়ে কিভাবে ভিন্নতর। দ্বিতীয়ত, পরীক্ষার দ্বারা ছাত্রের কোনো বিষয়ের দক্ষতার পুরোপুরি সঠিক মূল্যায়ন হয় না বলেই u_i এর ভেদ হয়। এখানে ছেলে-মেয়েদের পরীক্ষার ফলাফল মূল্যায়ন করার জন্য তিনটি চলক বিশ্লেষণ না করে একটি উপাদান f দ্বারাই ঐ মূল্যায়ন করা যায়। উপাদান বিশ্লেষণের মূল কাজই হলো তাই। এখানে f এর ভেদ x_i-গুলোর ভেদের অংশবিশেষ। এই f উপাদানের তথ্য নমুনা এককসমূহ হতে পাওয়া যায় না। এখানে উপাদান বিশ্লেষণের কাজ হলো x_i-সমূহের ভিত্তিতে f-কে চিহ্নিত করা।

৫.২ উপাদান বিশ্লেষণ মডেল (The Factor Analisis Model)

উপাদান বিশ্লেষণের মৌলিক মডেল হলো

$$X = \Lambda F + U \qquad (৫.২.১)$$

এখানে $X = p$-মাত্রার উপাত্ত ভেক্টর, $X' = (X_1, X_2, \ldots, X_p)$

$F = q$-মাত্রার সাধারণ উপাদান ভেক্টর, $F' = (f_1, f_2, \ldots, f_q)$

$\mathbf{U}$ = p-মাত্রার একক উপাদান (Unique factor)

ভেক্টর, $U = (e_1, e_2, \cdots, e_p)$

এই F ও U ভেক্টরের চলকসমূহের উপাত্ত সংগ্রহ করা যায় না।

$\Lambda = p \times q$ ম্যাট্রিক্স যার মানসমূহ হলো উপাদান ভর

অর্থাৎ
$$\Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1q} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2q} \\ \cdots & \cdots & \cdots & \cdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pq} \end{bmatrix}$$

এই λ_{ij} ($i = 1, 2, \cdots, p$; $j = 1, 2, \cdots, q$) হলো অজানা মান। অনুমান করতে হবে যে, একক উপাদানসমূহ $e_1, e_2, \cdots, e_p$ পরস্পর অনপেক্ষ এবং এগুলো সাধারণ উপাদানেরও অনপেক্ষ। অর্থাৎ

$$E = (UU') = \psi = \begin{bmatrix} \psi_1 & 0 & 0 & \cdots & 0 \\ 0 & \psi_2 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \\ 0 & 0 & 0 & \cdots & \psi p \end{bmatrix}$$

এবং $\mathrm{Cov}\,[\,U, F'\,) = 0$

আরো অনুমান করতে হবে যে,

$$E[F] = 0, \quad V[F] = I$$

উপরিউক্ত অনুমানের ভিত্তিতে মডেল ৫.২.১ হতে পাওয়া যায়

$$\Sigma = \Lambda \Lambda' + \psi \qquad (৫.২.২)$$

অবশ্য $\mathrm{Cov}\,(F) = \varphi = \begin{bmatrix} 1 & & & & \\ \varphi_{21} & 1 & & & \\ \varphi_{31} & \varphi_{32} & 1 & & \\ \cdots & \cdots & \cdots & \cdots & \\ \varphi_{q1} & \varphi_{q2} & \varphi_{q3} & \cdots & 1 \end{bmatrix}$

হলে $$\Sigma = \Lambda \Phi \Lambda' + \psi \qquad (৫.২.৩)$$

এখানে Σ হলো X ভেক্টরের সহ-ভেদাঙ্ক ম্যাট্রিক্স।

মডেল ৫.২.১-কে অন্যভাবেও লেখা যায়। তাহলো

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \lambda_{13} & \cdots & \lambda_{1q} \\ \lambda_{21} & \lambda_{22} & \lambda_{23} & \cdots & \lambda_{2q} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \lambda_{p1} & \lambda_{p2} & \lambda_{p3} & \cdots & \lambda_{pq} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_q \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_p \end{bmatrix} \qquad (৫.২.৪)$$

বা $$X_i = \sum_{j=1}^{q} \lambda_{ij} f_j + e_i \qquad (৫.২.৫)$$

$$i = 1, 2, \ldots, p$$

এখানে ৫.২.৪-এ উপস্থাপিত সমীকরণ গুচ্ছ হলো উপাদানের ধরন (factor pattern)। এই উপাদানসমূহ সম্পর্কে প্রাপ্ত চলকসমূহের ভিত্তিতে সিদ্ধান্ত নেয়া যায় এবং উপাদানসমূহকে চলকসমূহের ভিত্তিতে নিরূপণ করা যায়। ঐ নিরূপকগুলো হলো চলকসমূহের রৈখিক সমাবেশ। নিরূপণ পদ্ধতি ভিন্ন অনুচ্ছেদে আলোচনা করা হবে। তবে j-তম ($j = 1, 2, \cdots, q$) উপাদানের নিরূপকের সাধারণ আকার হলো

$$f_j = \sum_{i=1}^{p} W_{ij} X_i = W_{1j} X_1 + W_{2j} X_2 + \cdots + W_{pj} X_p \qquad (৫.২.৬)$$

প্রতিকৃতি ৫.২.৫ কিছুটা নির্ভরণ মডেলের ন্যায়। প্রতিটি চলককে উপাদানসমূহের রৈখিক সমাবেশ দ্বারা প্রকাশ করা হয়েছে। উদাহরণ হিসেবে একটি দেশের শিশু মৃত্যুর স্তর পর্যালোচনা করার জন্য উপাদান মডেল বিবেচনা করা যাক। ধরা যাক মডেল হলো

$$y = \alpha x_1 + \beta x_2 + \gamma x_3 + e \qquad (৫.২.৭)$$

এখানে y-কে বিবেচনা করা যাক, শিশু মৃত্যুর স্তর x_1 হলো শিক্ষার স্তরের পরিবর্তন, x_2 হলো অর্থনৈতিক অবস্থা, x_3 হলো স্বাস্থ্য সুযোগ এবং e হলো শিশু মৃত্যুর স্তরের ঐ অংশটুকু যা x_1, x_2 ও x_3 দ্বারা ব্যাখ্যা করা যায় না। এখানে মডেল ৫.২.৬ এবং নির্ভরণ মডেলের মধ্যে পার্থক্য হলো এই যে, এখানে আলোচিত চলকসমূহ x_1, x_2 ও x_3 এর উপাত্ত সরাসরি সংগ্রহ করা যায় না। এই

চলকগুলো হলো অন্য অনেকগুলো চলকের গুচ্ছ যা উপাদান হিসেবে পরিচিত। এই চলকগুলোর পরিমাপ করার জন্য মডেল ৫.২.৬ ব্যবহার করতে হয়। এখানে আলোচিত চলক x_1, x_2 ও x_3 হলো সাধারণ উপাদান, e হলো একক উপাদান। মডেল ৫.২.৭-কে নির্ভরণ মডেল বলা হয় যদি x_1, x_2 ও x_3 চলকত্রয় সরাসরি নমুনা একক হতে পরিমাপ করা যায়।

৫.৩ উপাদান ভর নিরূপণ (Estimation of factor loading)

মডেল ৫.২.৫ হতে লেখা যায়

$$V(X_i) = V\sum_{j=1}^{q} \lambda_{ij} f_j + V(e_i)$$

$$\sigma_i^2 = \sum_{j=1}^{q} \lambda_{ij}^2 + \psi_i \qquad \because V(f_j) = 1 \text{ এবং } Cov(f_j f_j') = 0$$

$$Cov(X_i, X_i') = Cov\left[\sum_j \lambda_{ij} f_j + e_i, \sum_j \lambda_{ij}' f_j + e_i'\right]$$

$$\sigma_{ii}' = \sum_{j=1}^{q} \lambda_{ij} \lambda_{ij}'$$

$$\therefore V(X) = \Sigma = \begin{bmatrix} \Sigma\lambda_{ij}^2 & \Sigma\lambda_{1j}\lambda_{2j} & \dots & \Sigma\lambda_{1j}\lambda_{pj} \\ \Sigma\lambda_{1j}\lambda_{2j} & \Sigma\lambda_{2j}^2 & \dots & \Sigma\lambda_{2j}\lambda_{pj} \\ \dots & \dots & \dots & \dots \\ \Sigma\lambda_{1j}\lambda_{pj} & \Sigma\lambda_{2j}\lambda_{pj} & \dots & \Sigma\lambda_{pj}^2 \end{bmatrix} + \begin{bmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \psi_p \end{bmatrix}$$

$$\Rightarrow \qquad \Sigma = \Lambda\Lambda' + \psi \qquad (৫.৩.১)$$

এখন $\Lambda\Lambda'$ এর কৌণিক মানকে লেখা যায়

$$\sigma_i^2 - \psi_i = \sum_j \lambda_{ij}^2$$

বা $$\sigma_i^2 - \sum_j \lambda_{ij}^2 + \psi_i = h_i^2 + \psi_i$$

এখানে $h_i^2 = \sum_{j=1}^{q} \lambda_{ij}^2$ কে বলা হয় কম্যুনালিটি (communality), σ_i^2 হলো

$V(X_i)$ যা সাধারণ উপাদানের মাধ্যমে অন্যান্য চলকের ভেদাঙ্কের অংশবিশেষ।
এখানে

$$\text{Cov}(X_i, f_j) = E[\lambda_{i1} f_1 + \lambda_{i2} f_2 + \cdots + \lambda_{iq} f_q, f_j]$$
$$= \lambda_{ij} \qquad \therefore E(X_i) = 0 \text{ এবং } E(f_i) = 0$$

এখান থেকে লেখা যায়

$$\text{Cov}(X F') = E[\Lambda F + U] F'$$
$$= \Lambda \qquad (৫.৩.২)$$

সুতরাং পরামান ম্যাট্রিক্স Λ হলো i-তম চলক ও j-তম উপাদানের সহভেদাঙ্ক। কিন্তু X_i গুলো আদর্শায়িত হলে Σ-ম্যাট্রিক্স সংশ্লেষাঙ্ক ম্যাট্রিক্স হয়। সেক্ষেত্রে λ_{ij} হলো i-তম চলক ও j-তম উপাদানের সংশ্লেষাঙ্ক। এ থেকে বলা যায় যে Λ এর মানসমূহ নিরূপণ করতে হবে

$$\sigma_i^2 = \sum_{j=1}^{q} \lambda_{ij}^2 + U_i \qquad (৫.৩.৩)$$

শর্তাধীনে।

সাধারণ উপাদান বিশ্লেষণ মডেলের ক্ষেত্রে পরামানের একক নিরূপক পাওয়া কষ্টকর। পরামান ম্যাট্রিক্সে সর্বমোট pq অজানা মান আছে—এগুলো হলো উপাদান ভর। উক্ত বিশ্লেষণের ক্ষেত্রে Σ-ম্যাট্রিক্স-এ $\frac{1}{2}p(p+1)$ ভিন্ন ভিন্ন ভেদাঙ্ক ও সহভেদাঙ্ক আছে। সুতরাং ৫.৩.১ হতে বলা যায় যে $\frac{1}{2}p(p+1)$ সমীকরণ আছে। কিন্তু নিরূপকসমূহের পরামান পেতে হলে সমীকরণগুলো আইডেনটি-ফায়েবল হতে হবে। সে কারণে পরামানের সংখ্যা সমীকরণের চেয়ে কম হতে হবে। অর্থাৎ $pq + p < \frac{1}{2}p(p+1)$ হওয়া উচিত। অথবা $q < \frac{1}{2}(p-1)$। এ থেকে বলা যায় p এর তুলনায় q বেশ ছোট হওয়া উচিত। অবশ্য q ছোট হলেই যে সমীকরণের সমাধান পাওয়া যাবে—এমন কথা নয়।

Lawley (1940, 1942, 1943) সর্বোত্তম সম্ভাব্যতা পদ্ধতি প্রয়োগ করে উপাদান ভরসমূহ (factor loadings) নিরূপণ করার প্রস্তাব করেছেন। তাঁর প্রস্তাব অনুযায়ী ধরা যাক $N = n+1$ আকারের অনপেক্ষ নমুনা উপাত্তের ভেক্টর আছে। ধরা যাক নমুনা উপাত্তসমূহ $N(\mu, \Sigma)$ হতে চয়ন করা হয়েছে এবং $r(\Sigma) = p$। উক্ত নমুনা হতে নমুনা সহ-ভেদাঙ্ক ম্যাট্রিক্স হলো S। যেহেতু S এর বিন্যাস হলো Wishatt বিন্যাস, এটির সম্ভাব্যতা ফাংশন হলো

$$f(S) = C|S|^{\frac{1}{2}(n-p-1)} |\Sigma|^{-\frac{1}{2}n} \exp(-\tfrac{1}{2}n \operatorname{tr} \Sigma^{-1} S) \qquad (৫.৩.৪)$$

এখানে C হলো S ব্যতীত রাশি। এই সম্ভাব্যতা ফাংশন-এর logarithm হলো

$$L(\Lambda, \psi) = \ln C + \tfrac{1}{2}(n-p-1)\ln |S| - \tfrac{1}{2}n \ln |\psi + \Lambda\Lambda'| - \tfrac{1}{2}n\,\mathrm{tr}[\psi + \Lambda\Lambda']^{-1}S$$

এখন Λ ও ψ এর $p(q+1)$ অজানা মান নিরূপণ করার জন্য সমীকরণ হবে

$$\frac{\partial L(\Lambda,\psi)}{\partial \psi_i} = 0 \quad \text{এবং} \quad \frac{\partial L(\Lambda,\psi)}{\partial \lambda_{ij}} = 0$$

এখানে
$$\frac{\partial L(\Lambda,\psi)}{\partial \psi_i} = -\frac{n}{2}\frac{1}{|\psi+\Lambda\Lambda'|}\frac{\partial |\psi+\Lambda\Lambda'|}{\partial \psi_i} + \frac{n}{2}\,\mathrm{tr}[\psi+\Lambda\Lambda']^{-1}\frac{\partial \psi}{\partial \psi_i}[\psi+\Lambda\Lambda^{-1}]^{-1}S$$

$$= -\frac{n}{2}\frac{1}{|\psi+\Lambda\Lambda'|}|\psi+\Lambda\Lambda'|_{ii} + \frac{n}{2}\,\mathrm{tr}[\psi+\Lambda\Lambda']^{-1}S[\psi+\Lambda\Lambda']^{-1}\frac{\partial\psi}{\partial\psi_i} \quad (৫.৩.৫)$$

এখানে $\partial\psi/\partial\psi_i$ হলো এমন একটি $(p \times p)$ কৌণিক মেট্রিক্স যার i-তম কৌণিক মান হলো 1 এবং অন্যান্য মান শূন্য। আবার $|\psi+\Lambda\Lambda'|_{ii}$ হলো $[\psi+\Lambda\Lambda']$ এর i-তম কৌণিক মানের কোফ্যাক্টর (cofactor)। এখন $\partial L(\Lambda,\psi)/\partial\psi_i = 0$ বসিয়ে পাওয়া যায়

$$\mathrm{diag}\{[\psi+\Lambda\Lambda']^{-1}(I - S[\psi+\Lambda\Lambda']^{-1})\} = 0$$

বা
$$\mathrm{diag}(\hat{\Sigma}^{-1}) = \mathrm{diag}(\hat{\Sigma}^{-1} S \hat{\Sigma}^{-1}) \quad (৫.৩.৬)$$

এখানে
$$\hat{\Sigma} = \hat{\psi} + \hat{\Lambda}\hat{\Lambda}' \quad (৫.৩.৭)$$

হলো উপাদান মডেল থেকে প্রাপ্ত নমুনা ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্স। এই $\hat{\Sigma}$ অবশ্য Σ-এর সর্বোত্তম সম্ভাব্য নিরূপক নয়। আবার,

$$\frac{\partial L(\Lambda,\psi)}{\partial \lambda_{ij}} = -\frac{n}{2}\frac{1}{|\psi+\Lambda\Lambda'|}\sum_{g=1}^{p}\sum_{h=1}^{p}|\psi+\Lambda\Lambda'|_{gh} \times \frac{\partial \sigma_{gh}}{\partial \lambda_{ij}} + \frac{n}{2}\,\mathrm{tr}[\psi+\Lambda\Lambda']^{-1}\frac{\partial\Sigma}{\partial\lambda_{ij}}[\psi+\Lambda\Lambda']^{-1}S$$

উপরের সমীকরণের ডানদিকের প্রথম রাশিটি হলো

$$-\frac{n}{2}\ \mathrm{tr}\ \Sigma^{-1}\frac{\partial\Sigma}{\partial\lambda_{ij}}$$

এখানে

$$\frac{\partial\Sigma}{\partial\lambda_{ij}}=\left[\frac{\partial\sigma_{gh}}{\partial\lambda_{ij}}\right]=\begin{bmatrix} 0 & \cdots & 0 & \lambda_{1j} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & \lambda_{i-1j} & 0 & \cdots & 0 \\ \lambda_{1j} & \cdots & \lambda_{i-1j} & 2\lambda_{ij} & \lambda_{i+1j} & \cdots & \lambda_{pj} \\ 0 & \cdots & 0 & \lambda_{i+1j} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & \lambda_{pj} & 0 & \cdots & 0 \end{bmatrix}$$

উপরের ম্যাট্রিক্সটি হলো প্রতিসম (symmetric) ম্যাট্রিক্স যার i-তম সারি এবং j-স্তম্ভ ছাড়া অন্যান্য মান শূন্য। ধরা যাক Σ^{-1} এর ij-তম মান হলো σ^{ij}, তাহলে

$$\begin{aligned}\mathrm{tr}\ \Sigma^{-1}\frac{\partial\Sigma}{\partial\lambda_{ij}} &= \sigma^{1i}\lambda_{1j}+\sigma^{2i}\lambda_{2j}+\cdots+\sigma^{ii}\lambda_{ij}+\cdots+2\sigma^{ii}\lambda_{ij}+\cdots \\ &\quad +\sigma^{ip}\lambda_{pj}+\cdots+\sigma^{ip}\lambda_{pj} \\ &= 2(\sigma^{i1}\lambda_{1j}+\cdots+\sigma^{ip}\lambda_{pj}) \\ &= 2\delta^{i}\tau_j\end{aligned}$$

এখানে δ^i হলো Σ^{-1} এর i-তম সারি এবং τ_j হলো Λ-এর j-তম স্তম্ভ। সুতরাং প্রথম রাশি $-\frac{n}{2}\mid n\mid\Sigma\mid$ এর pq বিযোজিত ফলাফলকে লেখা যায়

$$-n\,\Sigma^{-1}\Lambda$$

এবং দ্বিতীয় রাশি

$$\frac{n}{2}\ \mathrm{tr}\ \Sigma^{-1}\frac{\partial\Sigma}{\partial\lambda_{ij}}\Sigma^{-1}S=\frac{n}{2}\ \mathrm{tr}\ \Sigma^{-1}S\Sigma^{-1}\frac{\partial\Sigma}{\partial\lambda_{ij}}$$

এখন $\Sigma^{-1}S\Sigma^{-1}=Z=[Z_{gh}]$ ধরে, লেখা যায়

$$\mathrm{tr}\ Z\frac{\partial\Sigma}{\partial\lambda_{ij}}=2Z_i\tau_j \qquad (৫.৩.৮)$$

এখানে Z_i হলো Z-এর i-তম সারি। তাহলে

$$(\hat{\Sigma}^{-1} - \hat{\Sigma}^{-1} S \hat{\Sigma}^{-1}) \hat{\Lambda} = 0$$

বা
$$S \hat{\Sigma}^{-1} \hat{\Lambda} = \hat{\Lambda} \qquad (৫.৩.৯)$$

এখন ৫.৩.৬ ও ৫.৩.৯ সমীকরণ গুচ্ছদ্বয়কে সরল করে উপাদান ভরের নিরূপক পাওয়া যাবে। সমীকরণ ৫.৩.৬-এর বামদিক ও ডানদিক কৌণিক ম্যাট্রিক্স $\hat{\psi} = \hat{\Sigma} - \hat{\Lambda}\hat{\Lambda}'$ দ্বারা প্রাক ও পরা (post) গুণন করে পাওয়া যায়

$$\text{diag}(\hat{\Sigma} - 2\hat{\Lambda}\hat{\Lambda}' + \hat{\Lambda}\hat{\Lambda}\hat{\Sigma}^{-1}\hat{\Lambda}\hat{\Lambda}')$$
$$= \text{diag}(S - S\hat{\Sigma}^{-1}\hat{\Lambda}\hat{\Lambda}' - \hat{\Lambda}\hat{\Lambda}'\hat{\Sigma}^{-1}S + \hat{\Lambda}\hat{\Lambda}'\hat{\Sigma}^{-1}S\hat{\Sigma}^{-1}\hat{\Lambda}\hat{\Lambda}')$$

এখন ৫.৩.৯-এর মান বসিয়ে পাওয়া যায়

$$\text{diag}(\hat{\Sigma}) = \text{diag}(S) \qquad (৫.৩.১০)$$

সমীকরণ ৫.৩.৯-কে লেখা যায়

$$S\hat{\psi}^{-1}\hat{\Lambda}(I + \hat{\Lambda}'\hat{\psi}^{-1}\hat{\Lambda})^{-1} = \hat{\Lambda} \qquad (৫.৩.১১)$$

কারণ
$$(\hat{\psi} + \hat{\Lambda}\hat{\Lambda}')^{-1}\hat{\Lambda} = \hat{\psi}^{-1}\hat{\Lambda}(I + \hat{\Lambda}'\hat{\psi}^{-1}\hat{\Lambda})^{-1}$$

আবার, ৫.৩.১১-কে লেখা যায়

$$S\hat{\psi}^{-1}\hat{\Lambda} = \hat{\Lambda}(I + \hat{\Lambda}'\hat{\psi}^{-1}\hat{\Lambda}) \qquad (৫.৩.১২)$$

ধরা যাক $\hat{\Lambda}_0$ হলো $\hat{\Lambda}$-এর মান যা ৫.৩.১২ সমীকরণের ক্ষেত্রে সত্য। এখন $\hat{\Lambda}_0$ জানা বিবেচনা করে $\partial L(\Lambda\psi)/\partial\psi_1$ থেকে পাওয়া যায়

$$\text{diag}(\hat{\Lambda}_0\hat{\Lambda}_0' + \hat{\psi} - S) = 0 \qquad (৫.৩.১৩)$$

বা
$$\hat{\psi} = \text{diag}(S - \hat{\Lambda}_0\hat{\Lambda}_0') \qquad (৫.৩.১৪)$$

সমীকরণ ৫.৩.১৪ দেখতে সহজ হলেও এর সমাধান সরাসরি পাওয়া যায় না। এর সমাধান পাওয়ার জন্য ইটারেটিভ পদ্ধতি প্রয়োগ করতে হবে। এ সম্পর্কে বিস্তারিত জানার জন্য Lawley and Maxwell (1971) এবং Joreskog and Lawley (1968) আলোচনা করা যেতে পারে। এখানে $\hat{\psi}$-এর মান পাওয়ার জন্য একটি সংখ্যাভিত্তিক (numerical) সমাধান সংক্ষেপে আলোচনা করা হলো (Rao (1955), Maxwell (1964)।

সমীকরণ ৫.৩.১২-এর উভয় পাশে $\hat{\psi}^{-1/2}$ দ্বারা প্রাক-গুণ করে এবং সাজিয়ে পাওয়া যায়

$$[\hat{\psi}^{-1/2}(S-\hat{\psi})\hat{\psi}^{-1/2}]\hat{\Lambda}=\hat{\psi}^{-1/2}\hat{\Lambda}J \qquad (৫.৩.১৫)$$

এখানে $J=\hat{\Lambda}'\hat{\psi}^{-1}\hat{\Lambda}$ । $\hat{\Lambda}$-এর একক সমাধান পাওয়ার জন্য J কৌণিক ম্যাট্রিক্স হতে হবে। তাহলে $(S-\hat{\psi})\hat{\psi}^{-1}$ এবং $\hat{\psi}^{-1/2}(S-\hat{\psi})\hat{\psi}^{-1/2}$ এর প্রথম q আইগেন মান J ম্যাট্রিক্স-এর পর্যায়ক্রমিক মানসমূহ হবে $\hat{\psi}^{-1/2}\hat{\Lambda}$-এর i-তম স্তম্ভ হবে $\hat{\psi}^{-1/2}(S-\hat{\psi})\hat{\psi}^{-1/2}$-এর i-তম বৃহত্তম আইগেন মানের প্রাসঙ্গিক ভেক্টর। অবশ্য এই ভেক্টরের মান নির্ণয়ও সহজ নয়, কারণ $\hat{\psi}$ এর মান অজানা। $\hat{\psi}$-এর মান পেতে হবে $\hat{\psi}=\text{diag}(S-\hat{\Lambda}\hat{\Lambda}')$ থেকে। এ পর্যায়ে ইটারেটিভ পদ্ধতি প্রয়োগ করতে হবে এবং $\hat{\psi}$-এর একটি আনুমানিক মান ধরে নিতে হবে এবং ঐ মানের ভিত্তিতে আইগেন ভেক্টর পেতে হবে এবং $\hat{\Lambda}$-এর মান পেতে হবে।

ইটারেটিভ পদ্ধতির শুরুতে এক উপাদান (single factor) বিশিষ্ট মডেলের ক্ষেত্রে S-এর বৃহত্তম আইগেন মানের প্রাসঙ্গিক আইগেন ভেক্টরকে প্রাথমিক ভেক্টর বিবেচনা করা যাক। এই ভেক্টরের মানসমূহকে এমনভাবে মাপনী (scale) দ্বারা পরিবর্তন করতে হবে যেন মানসমূহের বর্গের যোগফল আইগেন মানের সমান হয়। তারপর ইটারেটিভ পদ্ধতি নিম্নরূপভাবে হবে।

১। S ম্যাট্রিক্স-এর আইগেন মান নির্ণয় কর। ধরা যাক l_{10} হলো ম্যাট্রিক্স-এর বৃহত্তম আইগেন মান এবং a_{10} হলো প্রাসঙ্গিক আইগেন ভেক্টর। আইগেন ভেক্টরের উপর এমনভাবে মাপনী পরিবর্তন করতে হবে যেন $a'_{10}a_{10}=l_{10}$ হয়।

২। ভেক্টর a_{10} ব্যবহার করে $\hat{\psi}_{10}=\text{diag}(S-a_{10}a'_{10})$ নির্ণয় করতে হবে। তারপর $\hat{\psi}_{10}$ ব্যবহার করে ম্যাট্রিক্স

$$\hat{\psi}_{10}^{-1/2}(S-\hat{\psi}_{10})\hat{\psi}_{10}^{-1/2} \qquad (৫.৩.১৬)$$

গঠন করতে হবে। ধরা যাক এই ম্যাট্রিক্স-এর বৃহত্তম আইগেন মান হলো l_{11}।

৩। l_{11} এর প্রাসঙ্গিক আইগেন ভেক্টর a_{11} এমনভাবে মাপনী পরিবর্তনের দ্বারা নিরূপণ করতে হবে যেন $a'_{11}a_{11}=l_{11}$ হয়।

৪। ভেক্টর a_{11} কে $\hat{\psi}_{10}^{1/2}$ দ্বারা প্রাক-গুণ করতে হবে। এই গুণফল হবে $\hat{\Lambda}_1$ ম্যাট্রিক্স-এর প্রথম স্তম্ভের প্রথম কাছাকাছি মান। ধরা যাক এই মান হলো v_{11}। এখানে $\hat{\Lambda}_1$ হলো এক স্তম্ভবিশিষ্ট ম্যাট্রিক্স।

৫। $\hat{\tau}_{111}$ ব্যবহার করে ম্যাট্রিক্স

$$\hat{\psi}_{111}=\text{diag}_i(S-\hat{\tau}_1\hat{\tau}_{11}')$$

নির্ণয় করতে হবে এবং এই ম্যাট্রিক্স-এর বৃহত্তম আইগেন মানের প্রাসঙ্গিক আইগেন ভেক্টর a_{12} নির্ণয় করে তার উপর মাপনী পরিবর্তন করতে হবে যেন $a_{12}'a_{12}=l_{12}$ হয়। এখানে l_{12} হলো আইগেন মান। তারপর উপরে আলোচিত নিয়মে $\hat{\Lambda}_1$ এর দ্বিতীয় কাছাকাছি মান নিরূপণ করতে হবে। ঐ মান ধরা যাক $\hat{\tau}_{12}$।

৬। এভাবে ইটারেটিভ পদ্ধতি চলতে থাকবে এবং যতক্ষণ পর্যন্ত $\hat{\tau}_{11}$ ও $\hat{\tau}_{11+1}$ এর মধ্যে একটি পূর্ব নির্ধারিত মানের বেশি পার্থক্য না হয় ততক্ষণ ইটারেশন চলবে। এখানে i দ্বারা i-তম ইটারেশন বুঝানো হয়েছে। এভাবে ইটারেশন করা হলে প্রাপ্ত $\hat{\Lambda}_1$ এক উপাদান প্রতিকৃতির উপাদান ভরের সর্বোত্তম সম্ভাব্য নিরূপক হবে।

উপরিউক্ত নিয়ম ইটারেশন করা হলে তা q সংখ্যক উপাদান বিশ্লেষণের ক্ষেত্রে যথাযথ হবে না। সেক্ষেত্রে প্রধান উপাদান বিশ্লেষণ (Principal component analysis) হতে প্রাপ্ত উপাদান ভর এবং উপাদান বিশ্লেষণ (Factor analysis) হতে প্রাপ্ত উপাদান ভর এক হবে না। q সংখ্যক উপাদানবিশিষ্ট উপাদান বিশ্লেষণের ক্ষেত্রে $\hat{\Lambda}_1$ নিরূপণ করার পর অবশিষ্ট ম্যাট্রিক্স (Residual matrix) নির্ণয় করার প্রস্তাব করেছেন Maxwell। এই অবশিষ্ট ম্যাট্রিক্স হলো

$$S_1=S-\hat{\Lambda}_1\hat{\Lambda}_1'$$

ধরা যাক S_1-এর বৃহত্তম আইগেন মানের প্রাসঙ্গিক আইগেন ভেক্টর হলো a_{20}। একে মাপনী পরিবর্তন দ্বারা এমন করতে হবে যেন $a_{20}'a_{20}=l_{20}$ হয়, এখানে l_{20} হলো S_1-এর বৃহত্তম আইগেন মান। এখন এক উপাদান প্রতিকৃতির জন্য প্রাপ্ত সমাধান $\hat{\Lambda}_1$ ও a_{20}-এর সমন্বয়ে একটি $(p\times 2)$ ম্যাট্রিক্স

$$\hat{\Lambda}_{20}=[\hat{\Lambda}_1 a_{20}]$$

নির্ণয় করতে হবে। $\hat{\Lambda}_{20}$ ব্যবহার করে

$$\hat{\psi}_{20}=\text{diag}\,(S-\hat{\Lambda}_{20}\hat{\Lambda}_{20}')$$

এবং $$\hat{\psi}_{20}^{-1/2}(S-\hat{\psi}_{20})\hat{\psi}_{20}^{-1/2}$$

নির্ণয় করতে হবে। শেষোক্ত ম্যাট্রিক্স-এর সবচেয়ে বড় আইগেন মানদ্বয়ের প্রাসঙ্গিক আইগেন ভেক্টর নির্ণয় করে তাদের উপর মাপনী পরিবর্তন করে তাদেরকে $(p \times 2)$ ম্যাট্রিক্স $(a_{11}\ a_{24})$ আকারে সাজাতে হবে। মনে রাখতে হবে যে, এক উপাদানী প্রতিকৃতির ক্ষেত্রে প্রাপ্ত a_{11} এবং $(p \times 2)$-এর a_{11} এক হবে না। এখন দুই উপাদানী প্রতিকৃতির ক্ষেত্রে প্রথম কাছাকাছি উপাদান ভরের নিরূপক $\hat{\Lambda}_{21}$ পাওয়ার জন্য $(a_{11}\ a_{21})$-কে $\hat{\psi}_{10}^{-1/2}$ দ্বারা প্রাকগুণ করতে হবে। এভাবে ইটারেশন পদ্ধতি চলতে থাকবে যতক্ষণ না $\hat{\Lambda}_{21}$ একটি নির্দিষ্ট সঠিকতাসহ $\hat{\Lambda}_2$-এর মান দিবে।

q-উপাদানবিশিষ্ট উপাদান বিশ্লেষণের ক্ষেত্রে অনুরূপ পদ্ধতি প্রয়োগ করতে হবে। এক্ষেত্রে

$$S_{q-1} = S - \hat{\Lambda}_{q-1} \hat{\Lambda}'_{q-1}$$

অবশিষ্ট ম্যাট্রিক্স-এর সবচেয়ে বড় আইগেন মানসমূহ এবং এদের প্রাসঙ্গিক আইগেন ভেক্টরসমূহ নির্ণয় করতে হবে।

আগেই উল্লেখ করা হয়েছে যে, Λ এর মান $\Sigma = \Lambda\Lambda' + \psi$ হতে নিরূপণ করা যায়। এখানে সমীকরণে Λ এবং ψ-এর মধ্যে $pq + p$ পরামান আছে এবং Σ-এর মধ্যে আছে $\frac{1}{2}p(p+1)$ পরামান। লক্ষ্য করা যাচ্ছে যে, এখানে সমীকরণের সংখ্যা এবং পরামানের সংখ্যা সমান নয়। এক্ষেত্রে সমীকরণের সমাধান পাওয়ার জন্য $\Lambda'\psi^{-1}\Lambda$ কৌণিক হবে বা $\Lambda'D^{-1}\Lambda$ কৌণিক হবে শর্ত আরোপ করা যেতে পারে, এখানে $D = (\sigma_1^2\sigma_2^2\cdots\sigma_p^2)$। এই শর্তের কারণে উপাদান প্রতিকৃতির $\frac{1}{2}q(q-1)$ পরামান শূন্য বিবেচনা করা হয়। সমীকরণের সমাধান পাওয়ার জন্য Σ ও উপাদান প্রতিকৃতির পরামানসমূহের পার্থক্য বিবেচনা করা প্রয়োজন। ধরা যাক এই পার্থক্য হলো

$$\begin{aligned} s &= \tfrac{1}{2}p(p+1) - \{pq + p \quad \tfrac{1}{2}(q-1)q\} \\ &= \tfrac{1}{2}(p-q)^2 - \tfrac{1}{2}(p+q) \end{aligned} \qquad (৫.৩.১৭)$$

এখানে, $s < 0$ হলে, সমীকরণের সংখ্যার চেয়ে পরামানের সংখ্যা বেশি হলে, একক সমাধান পাওয়া যাবে না। এরূপ ক্ষেত্রে উপাদান প্রতিকৃতি সঠিকভাবে সংজ্ঞায়ন করা নয় বলে বিবেচিত হবে। যদি $s = 0$ হয়, অর্থাৎ সমীকরণের সংখ্যা ও পরামানের সংখ্যা সমান হয়, তাহলে Λ-ও ψ-এর সঠিক সমাধান পাওয়া যাবে। আবার, $s > 0$ হলে Λ-ও ψ-এর কাছাকাছি সমাধান পাওয়া যাবে।

সর্বোত্তম সম্ভাব্য নিরূপকের একটি গুরুত্বপূর্ণ ধর্ম হলো যে i-তম চলকের ভেদাঙ্ক σ_i^2-এর নিরূপক হলো

$$\sigma_i^2 = \sum_{j=1}^{q} \hat{\lambda}_{ij}^2 + \hat{\psi}_i^2$$

৫.৪ উপাদান প্রতিকৃতি মাপনী দ্বারা পরিবর্তিত হয় না (Factor Model is Scale Invariance)

ধরা যাক X চলক ম্যাট্রিক্স-এর ক্ষেত্রে উপাদান প্রতিকৃতি হলো

$$X = \Lambda_x F + U$$

যেখানে $$V(X) = \Sigma = \Lambda_x \Lambda_x' + \psi_x$$

এখন X ম্যাট্রিক্সকে Z দ্বারা মাপনী পরিবর্তন করা যাক। ধরা যাক $Z = CX$, যেখানে $C = \text{diag}\ (C_i)$। তাহলে

$$Z = CX = C\Lambda_x F + CU$$

$$V(z) = C\Sigma C' = C\Lambda_x C' + C\psi_x C'$$

বা $$C\Sigma C' = C\Lambda_x \Lambda_x' C + C\psi_x C \qquad [\because C = C']$$

এখানে উপাদান ভর ম্যাট্রিক্স হলো $C\Lambda_x$ এবং একক উপাদান ভেক্টর ভেদাঙ্ক হলো

$$\psi_z = C\psi_x C = \text{diag}\ (C_i^2 \psi_i^2)$$

লক্ষ্য করা যাচ্ছে যে, মাপনী দ্বারা পরিবর্তিত চলক Z-এর উপাদান ভর ম্যাট্রিক্স আদি চলক X-এর উপাদান ভরের C গুণ $(C\Lambda_x)$, যেখানে C হলো মাপনী ম্যাট্রিক্স। কাজেই উপাদান বিশ্লেষণ মাপনী দ্বারা পরিবর্তিত হয় না।

৫.৫ সংশ্লেষাঙ্ক ম্যাট্রিক্স R হতে উপাদান ভর নিরূপণ (Estimation of Factor Loadings from Correlation Matrix R)

উপাদান প্রতিকৃতি বিবেচনা করা হয়েছে

$$X = \Lambda F + U \qquad (৫.৫.১)$$

ধরা যাক চলকসমূহ আদর্শায়িত করা হয়েছে। তাহলে আদর্শায়িত চলক হবে

$$Y = H \times D_S^{-1/2}, \text{ এখানে}$$

$$D_S = \text{diag}(s_1^2, s_2^2, \ldots, s_p^2), \quad H = I - \frac{1}{n} 11', \quad 1 = [\,1 \;\; 1 \cdots 1\,]'$$

এখানে আদি চলকসমূহকে আদর্শায়িত করার কারণে S ম্যাট্রিক্স R ম্যাট্রিক্স-এ পরিবর্তিত হয়েছে। আগেই আলোচনা করা হয়েছে উপাদান প্রতিকৃতি মাপনী দ্বারা পরিবর্তিত হয় না। আদর্শায়িত করার কারণে আদি চলকসমূহের উপরে মাপনী-এর প্রভাবই থাকে। ফলে নতুন চলক ম্যাট্রিক্স Y-এর ক্ষেত্রে

$$\hat{\Lambda}_y = D_S^{-1/2} \hat{\Lambda}_x \text{ এবং } \hat{\psi}_y = D_S^{-1} \hat{\psi}_x$$

সুতরাং সংশ্লেষাঙ্ক ম্যাট্রিক্সকে লেখা যায়

$$\mathbf{R} = \hat{\Lambda}_{\mathbf{y}} \hat{\Lambda}_{\mathbf{y}}' + \hat{\psi}_{\mathbf{y}} \qquad (৫.৫.২)$$

কাজেই $$\hat{\psi}_{yi}^2 = 1 - \sum_{i=1}^{q} \hat{\lambda}_{yi}^2 \qquad (i = 1, 2, \ldots, p)$$

এবং $\psi_{\mathbf{y}}$ আর প্রতিকৃতির পরামান থাকছে না। অবশ্য S অপেক্ষা R-এর মধ্যে P পরামান কম থাকে। তবে এক্ষেত্রেও s এর মান ৫.৩.১৭ দ্বারা নির্ণয় করা যায় এবং মাপনী পরামানের নিরূপক পাওয়ার জন্য $\hat{\sigma}_i^2 = s_i^2$ $(i = 1, 2, \ldots, p)$ ব্যবহার করা যায়।

উপাদান বিশ্লেষণ করার জন্য R ম্যাট্রিক্স ব্যবহার করার আগে নাস্তিকল্পনা $H_0 : P = I$ যাচাই করা প্রয়োজন। কারণ উপাদান বিশ্লেষণের একটি উদ্দেশ্য হলো এমন কিছু উপাদান নির্ণয় করা যেগুলো চলকসমূহের সংশ্লেষণ ব্যাখ্যা করতে পারে। সেক্ষেত্রে চলকসমূহ অসংশ্লেষিত হলে এদের মধ্যে সাধারণ উপাদান (common factor) থাকার কথা নয়। উক্ত নাস্তিকল্পনা যাচাই করার জন্য যাচাই তথ্যজমান হলো

$$-2 \log \lambda = -n \log |R| \qquad (৫.৫.৩)$$

এখানে P হলো গণসমষ্টি সংশ্লেষাঙ্ক ম্যাট্রিক্স, R হলো নমুনা সংশ্লেষাঙ্ক ম্যাট্রিক্স। এই $-2 \log \lambda$ অভিসারীভাবে $\frac{1}{2} p(p-1)$ স্বাধীনতার মাত্রাসহ কাই-বর্গ বিন্যাস অনুসরণ করে। Box (1949) দেখিয়েছেন যে n এর পরিবর্তে $n - \frac{1}{6}(2p+11)$ ব্যবহার করা হলে $-2 \log \lambda$ এর বিন্যাস χ^2 বিন্যাসের বেশি কাছাকাছি হয়।

চলকসমূহের সম্পর্কের মাত্রা পর্যালোচনা করার জন্য আংশিক সংশ্লেষাঙ্কও নির্ণয় করা যেতে পারে। যদি চলকসমূহের মধ্যে সাধারণ উপাদান থাকে, তাহলে অন্য চলকের বৈধিক প্রভাব দূরীভূত করা হলে যে কোনো চলকদ্বয়ের আংশিক সংশ্লেষাঙ্ক ছোট হবে। আংশিক সংশ্লেষাঙ্ক একক উপাদানসমূহের মধ্যে যে সংশ্লেষাঙ্ক আছে তার নিরূপক এবং উপাদান বিশ্লেষণের অনুমান বহাল থাকলে এই সংশ্লেষাঙ্ক শূন্য-এর কাছাকাছি হবে।

উপাদান বিশ্লেষণ শুরু করার পূর্বে প্রাপ্ত নমুনা ঐ বিশ্লেষণের জন্য যথোপযুক্ত কিনা তা নির্ধারণ করার জন্য Kaiser-Meyer-Olkin (KMO) পরিমাপ নির্ণয় করা যায়। তাঁদের মতে

$$KMO = \frac{\sum\sum_{i \neq j} r_{ij}^2}{\sum\sum_{i \neq j} r_{ij}^2 + \sum\sum a_{ij}^2}$$

এখানে r_{ij} ($i \neq j = 1, 2, \cdots, p$) হলো i-তম ও j-তম চলকের সংশ্লেষাঙ্ক, a_{ij} হলো i-তম ও j-তম চলকের আংশিক সংশ্লেষাঙ্ক। $\Sigma\Sigma a_{ij}^2$ খুব ছোট হলে KMO = 1 এবং $\Sigma\Sigma a_{ij}$ ছোট হওয়ার অর্থই হলো a_{ij} গুলো খুব ছোট হওয়া। সেক্ষেত্রে উপাদান বিশ্লেষণ অর্থবহ হতে পারে। কিন্তু KMO-এর মান ছোট হলে উপাদান বিশ্লেষণ অর্থবহ হবে না, কারণ চলকগুলোর জোড়ায় জোড়ায় যে সংশ্লেষণ তা অন্য কোনো চলক দ্বারা ব্যাখ্যা করা যাবে না। Kaiser (1974) উল্লেখ করেছেন যে, KMO 0.90 হলে উপাদান বিশ্লেষণ খুবই অর্থবহ, 0.80 হলে ভাল, 0.70 হলে মোটামুটি, 0.60 হলে উপাদান বিশ্লেষণ করা যেতে পারে, 0.50 হলে বিশ্লেষণ করা ভাল নয় এবং 0.50 এর ছোট হলে বিশ্লেষণ গ্রহণযোগ্য নয়।

কোনো একটি চলক উপাদান বিশ্লেষণে অন্তর্ভুক্ত হবে কিনা সে সম্পর্কে সিদ্ধান্ত নেয়ার জন্য নমুনা পর্যাপ্ততার পরিমাপ (Measure of simpling adequacy, MSA) নির্ণয় করা যায়। i-তম চলকের ($i = 1, 2, \cdots, p$) জন্য MSA_i হলো

$$MSA_i = \frac{\sum\sum_{j \neq i} r_{ij}^2}{\sum_{j \neq i} r_{ij}^2 + \sum_{j \neq i} a_{ij}^2}$$

এই MSA_i এর মান ছোট হলে i-তম চলককে বিশ্লেষণ হতে বাদ দেয়া যুক্তিসঙ্গত।

কোনো চলককে উপাদান বিশ্লেষণ হতে বাদ দেয়ার জন্য ঐ চলকের সাথে অন্যান্য চলকের বহু সংশ্লেষাঙ্কের বর্গও (R^2) পর্যালোচনা করা যেতে পারে। কোনো চলকের ক্ষেত্রে R^2 ছোট হলে ঐ চলককে উপাদান বিশ্লেষণ হতে বাদ দেয়া যুক্তিসঙ্গত।

৫.৬ উপাদান নির্ধারণ (Factor Extraction)

উপাদান বিশ্লেষণের ক্ষেত্রে উপাদানের সংখ্যা নির্ধারণ এবং সাধারণ উপাদানের নিরূপক পাওয়া একটি সমস্যা। নিরূপক পাওয়ার অনেক পদ্ধতি আছে। উপাদানের

সংখ্যার সঠিকতা যাচাই (৫.৭ অনুচ্ছেদ)-এর পরিপ্রেক্ষিতে পদ্ধতিগুলো বিভিন্ন। বিভিন্ন পদ্ধতির মধ্যে একটি হলো সর্বোত্তম সম্ভাব্যতা পদ্ধতি যা (৫.৪) অনুচ্ছেদে আলোচনা করা হয়েছে। আরো আরো পদ্ধতির মধ্যে উল্লেখযোগ্য পদ্ধতিসমূহ হলো (ক) প্রধান উপাদান বিশ্লেষণ (principal component analysis), (খ) প্রধান উপাদান পদ্ধতি (principal factor method), (গ) অভরারোপিত ন্যূনতম বর্গ পদ্ধতি (unweighted least squares) এবং (ঘ) জেনারালাইজড্ ন্যূনতম বর্গ পদ্ধতি (generalised least squares)। অভরারোপিত ন্যূনতম বর্গ পদ্ধতি একটি নির্দিষ্ট সংখ্যক উপাদানের জন্য উপাদান প্যাটার্ন ম্যাট্রিক্স (factor pattern matrix) দিয়ে থাকে। এই ম্যাট্রিক্স প্রাপ্ত সংশ্লেষাঙ্ক ম্যাট্রিক্স ও বিশ্লেষিত সংশ্লেষাঙ্ক ম্যাট্রিক্স-এর বর্গের পার্থক্যের যোগফলকে ন্যূনতম করে। জেনারেলাইজড্ ন্যূনতম বর্গ পদ্ধতিও একই কাজ করে। কিন্তু এক্ষেত্রে চলকের এককতা (uniqueness) দ্বারা সংশ্লেষণকে উল্টোভাবে ভরারোপিত করা হয়।

প্রধান উপাদান বিশ্লেষণ (principal component analysis)-এর মাধ্যমেও উপাদান নির্ণয় করা যায়। এ সম্পর্কে বিস্তারিত আলোচনা চতুর্থ অধ্যায়ে করা হয়েছে। অবশ্য প্রধান উপাদান বিশ্লেষণ এবং উপাদান বিশ্লেষণ-এর মধ্যে পার্থক্য আছে। তবে উপাদান বিশ্লেষণের মাধ্যমে চলকসমূহের কতগুলো অসংশ্লেষিত রৈখিক সমাবেশ পেতে হলে প্রধান উপাদান বিশ্লেষণ পদ্ধতি প্রয়োগ করা যায়। এই বিশ্লেষণের মাধ্যমে উপাদানের সংখ্যা নির্ধারণ করার বিষয়েও চতুর্থ অধ্যায়ে আলোচনা করা হয়েছে।

প্রধান উপাদান পদ্ধতি (Principal Factor Method)

ধরা যাক উপাত্ত ম্যাট্রিক্স X-কে সংশ্লেষাঙ্ক ম্যাট্রিক্স R দ্বারা সংক্ষেপে প্রকাশ করা হয়েছে। অর্থাৎ উপাদান প্রতিকৃতির Λ ও Ψ এর মান আদর্শায়িত চলকসমূহ হতে নিরূপণ করতে হবে। এই পদ্ধতিতে উপাদান এমনভাবে নির্ধারণ করতে হবে যেন প্রতিটি উপাদান চলকসমূহের ভেদের বৃহত্তর অংশ ব্যাখ্যা করে। এই পদ্ধতিতে যেহেতু R ম্যাট্রিক্স ব্যবহার করা হয়, সে কারণে এই পদ্ধতি হলো $s=\frac{1}{8}(p-q)^2 - \frac{1}{2}(p+q) > 0$ এর ক্ষেত্রে পরামান নির্ণয় করার পদ্ধতি। এটি প্রধান উপাদান বিশ্লেষণ পদ্ধতির অনুরূপ। পার্থক্য হলো, এই পদ্ধতিতে R ম্যাট্রিক্স-এর কৌণিক মানসমূহের পরিবর্তে নিরূপিত কম্যুনালিটি-এর মান বসাতে হবে।

ধরা যাক $\tilde{h}_i$ হলো কম্যুনালিটি h_i^2 এর নিরূপক, যেখানে

$$\tilde{h}_i^2 = \frac{r_{ij}\, r_{ik}}{r_{jk}}$$

এখানে X_j এবং X_k হলো এমন দুটি চলক যেগুলো X_i এর সাথে সবচেয়ে বেশি সংশ্লেষিত। $\tilde{h_i}^2$ এর অন্য মানও নেয়া যায়। যেমন,

(ক) i-তম চলকের সাথে অন্যান্য চলকের বহু সংশ্লেষাঙ্কের বর্গ (R_i^2)।

(খ) i-তম চলকের সাথে অন্য কোনো একটি চলকের বৃহত্তম সংশ্লেষাঙ্ক।

আগেই লক্ষ্য করা গেছে যে, সকল চলকের ভেদাঙ্কের পরিমাণে উপাদান F_j এর ভেদাঙ্কের পরিমাণ হলো

$$V_j = \sum_{i=1}^{p} \lambda_{ij}^2 = \tau_j' \tau_j$$

এখানে τ_j হলো Λ-এর j-তম স্তম্ভ। কাজেই চলকসমূহের মোট ভেদাঙ্কে উপাদান-সমূহের ভেদাঙ্কের পরিমাণ হলো

$$V = \sum_{j=1}^{q} V_j$$

সুতরাং, মোট ভেদাঙ্কে j-তম উপাদানের ভেদাঙ্কের হার হলো $V_c = V_j/V$। যেহেতু চলকসমূহ আদর্শায়িত করা তাদের মোট ভেদাঙ্কের পরিমাণ হলো চলকের সংখ্যা P-এর সমান। চতুর্থ অধ্যায়ে লক্ষ্য করা গেছে যে উৎপাদক F_j এর ভেদাঙ্ক হলো λ_j। সুতরাং, j-তম চলকের ক্ষেত্রে $V_c = \lambda_j/\sum\lambda_j$। এখানে V হলো মোট কম্যুনালিটি। প্রধান উপাদান পদ্ধতি শুরু করা হয় এমনভাবে যেখানে F_1 এর জন্য সহগসমূহ $\lambda_{11}, \lambda_{21}, \ldots, \lambda_{p1}$ এমনভাবে নির্ণয় করতে হয় যেন V-এর মধ্যে F_1 এর ভেদাঙ্কের পরিমাণ বেশি হয়। এখানে শর্ত হলো p কম্যুনালিটি থাকবে এবং $p(p-1)/2$ সংশ্লেষাঙ্ক থাকবে।

Harman (1976) দেখিয়েছেন যে $\lambda_{11}, \lambda_{21}, \ldots, \lambda_{p1}$ এর মান পাওয়ার জন্য সংকুচিত সংশ্লেষাঙ্ক ম্যাট্রিক্স (reduced correlation matrix) $R^* = R - \tilde{\psi}$ এর আইগেন মান ও আইগেন ভেক্টর নির্ণয় করতে হয়। এখানে $\tilde{h_i}^2 = 1 - \tilde{\psi_i}^2$। এক্ষেত্রে R^* এর বৃহত্তম আইগেন মান হলো $V_1 = \tau_1' \tau$। ধরা যাক δ_1 হলো R^* এর বৃহত্তম আইগেন মান এবং প্রাসঙ্গিক নরমালাইজড্ আইগেন ভেক্টর হলো Y_1। তাহলে $\tau_1 = \sqrt{\delta_1} Y_1$।

একই পদ্ধতিতে দ্বিতীয় উপাদানের ভর τ_2 নিরূপণ করতে হবে। সেক্ষেত্রে R^*-কে আবার সমন্বয় করতে হবে। প্রথম সমন্বয় করা সংকুচিত সংশ্লেষাঙ্ক

ম্যাট্রিক্স হলো

$$R_1^* = R^* - \tau_1 \tau_1'$$

এখন $V_2 = \tau_2' \tau_2$ সর্বোচ্চ হয় এবং

$$R_1^* = \sum_{j=2}^{p} \tau_j \tau_j'$$

শর্তে τ_2 নির্ণয় করতে হবে। ধরা যাক R_1^* এর বৃহত্তম আইগেন মান হলো δ_2 এবং এর প্রাসঙ্গিক নরমালাইজড্ ভেক্টর হলো Y_2, তাহলে $\tau_2 = \sqrt{\delta_2} Y_2$। এখানে V_2 হলো δ_2। অনুরূপভাবে $\tau_3, \tau_4, \ldots, \tau_q$ নির্ণয় করতে হবে।

এই পদ্ধতিতে উপাদান নির্ণয় করার ক্ষেত্রে q এর মান কত হবে তা পূর্ব নির্ধারিত থাকে না। তবে প্রতি পর্যায়ে একটি করে উপাদান বাদ দিতে হয় এবং সংকুচিত সংশ্লেষাঙ্ক ম্যাট্রিক্সকে সমন্বয় করতে হয়। ধরা যাক m উপাদানের জন্য $\tau_1, \tau_2, \ldots, \tau_m$ নির্ণয় করা হয়েছে। সুতরাং, m উপাদান বাদ দিয়ে সমন্বিত সংকুচিত সংশ্লেষাঙ্ক ম্যাট্রিক্স হবে

$$R_m^* = R^* - \sum_{j=1}^{m} \tau_j \tau_j'$$

এক্ষেত্রে R_m^* এর মানসমূহ খুব ছোট হলে পুনরায় উপাদান নির্ণয় করার তেমন অর্থবহ নয়। অর্থাৎ এখানে m উপাদানই যথেষ্ট।

R^* এর কৌণিক মানগুলো কম্যুনালিটি দ্বারা পরিবর্তন দ্বারা পরিবর্তন করা হয় বলে অধিকাংশ ক্ষেত্রেই ঐ কৌণিক মানসমূহ 1 এর ছোট হয়। সেক্ষেত্রে R^* ধনাত্মক সেমিডেফিনিট থাকে না বলে কিছু কিছু আইগেন মান ঋণাত্মক হয় এবং প্রাসঙ্গিক আইগেন ভেক্টর কল্পিত মান নেয়। ঐরূপ হলে উক্ত আইগেন মান ও আইগেন ভেক্টর বাদ দিতে হয়। এরূপ ক্ষেত্রে উপাদানের সংখ্যা এমনভাবে নির্ধারণ করতে হবে যেন আইগেন মানসমূহের যোগফল মোট কম্যুনালিটির সমান হয়।

প্রধান উপাদান পদ্ধতিতে কম্যুনালিটিগুলির পরিবর্তে 1 ব্যবহার করা হলে বা 1 এর কাছাকাছি মান ব্যবহার করা হলে এবং q যদি p এর সমান হয়, তাহলে প্রধান উপাদান পদ্ধতি (principal factor method) এবং প্রধান উপাদান বিশ্লেষণ (principal component analysis) একই রকম ফলাফল দিবে। এখানে একটি বিষয় উল্লেখযোগ্য যে, প্রধান উপাদান পদ্ধতিতে নির্ধারিত উপাদান মাপনীর প্রভাব মুক্ত নয় (not scale invariant)।

৫.৭ উপাদান সংখ্যার পর্যাপ্ততা যাচাই (Test of Sufficiency of Factor Number)

ধরা যাক একটি উপাত্ত ম্যাট্রিক্স $\mathbf{X}$ এর মোট ভেদাঙ্ক q সাধারণ উপাদান দ্বারা প্রকাশ করা যায় এবং এই q উপাদানই উপাত্তের ভেদ প্রকাশ করার জন্য যথেষ্ট। এক্ষেত্রে উপাদান প্রতিকৃতির জন্য q-সংখ্যক উপাদানের পর্যাপ্ততা যাচাই করা যেতে পারে। এই যাচাই-এর জন্য নাস্তিকল্পনা হলো, H_0 : q সাধারণ উপাদানই পর্যাপ্ত। অর্থাৎ

$$H_0 : \Sigma = \Lambda \Lambda' + \psi \qquad (৫.৭.১)$$

এখানে Λ হলো $(p \times q)$ মাত্রার ম্যাট্রিক্স।

বিকল্প কল্পনা হলো

H_A : Σ হলো একটি $(p \times p)$ মাত্রার ধনাত্মক ডেফিনিট ম্যাট্রিক্স। উক্ত নাস্তিকল্পনা যাচাই করার জন্য সম্ভাব্যতা অনুপাত যাচাই [LRT] তথ্যজমান হলো

$$-2 \log \lambda = np(a - \log g - 1) \qquad (৫.৭.২)$$

এখানে a এবং g হলো $(\hat{\Sigma}^{-1} S)$ এর আইগেন মানসমূহের যথাক্রমে গাণিতিক ও জ্যামিতিক গড়, $\hat{\Sigma} = \hat{\Lambda} \hat{\Lambda}' + \hat{\psi}$, S হলো নমুনা ভেদাঙ্ক ম্যাট্রিক্স। এই $-2 \log \lambda$ অভিসারীভাবে $\frac{1}{2}[(p-q)^2 - (p+q)]$ স্বাধীনতার মাত্রাবিশিষ্ট কাইবর্গ বিন্যাস অনুসরণ করে। Bartlett (1954) দেখিয়েছেন যে n এর পরিবর্তে $[(n-1) - 1/6(2p+5) - \frac{2}{3}q)]$ ব্যবহার করা হলে $-2 \log \lambda$ এর বিন্যাস χ^2 বিন্যাসের বেশি কাছাকাছি হয়।

বাস্তবে q এর সংখ্যা কত হবে তা সঠিকভাবে নির্ধারণ করা সহজ নয়। অবশ্য উপাদানের সংখ্যা নির্ধারণের জন্য λ নির্দেশক প্রয়োগ করার বিষয় উল্লেখ করা আছে। যতগুলো λ এর মান 1 বা তার বেশি হবে ততগুলো উপাদান নির্ধারণ করার প্রস্তাব অনেকে করেছেন। আবার, যে সমস্ত উপাদান দ্বারা মোট ভেদের 90% বা বেশি ভেদ ব্যাখ্যা করা যায় সে সব উপাদান বিশ্লেষণে অন্তর্ভুক্ত করার প্রস্তাবও কেহ কেহ করেছেন। তবে ঐ সমস্ত প্রস্তাব থেকে প্রতিকৃতির পর্যাপ্ততা সম্পর্কে সিদ্ধান্ত নেয়া যায় না। কাজেই পর্যাপ্ততা যাচাই করাই যুক্তিসঙ্গত। অবশ্য উপাদান নির্ধারণ সর্বোত্তম সম্ভাব্য পদ্ধতির মধ্যে করা হলেই পর্যাপ্ততা যাচাই যুক্তিসঙ্গত।

উপরে আলোচিত পর্যাপ্ততা যাচাই-এর ক্ষেত্রে $q = 0$ হলে বুঝতে হবে যে চলকসমূহ অনপেক্ষ এবং সেক্ষেত্রে Σ-এর সর্বোত্তম সম্ভাব্য নিরূপক হলো $\hat{\Sigma} = \text{diag}(S)$ এবং $\hat{\Sigma}^{-1} S$ এর আইগেন মানসমূহ $\hat{\Sigma}^{-1/2} S \hat{\Sigma}^{-1/2} = R$ এর আইগেন মানসমূহের সমান। সেক্ষেত্রে $-2 \log \lambda = -n \log |R|$।

আগেই উল্লেখ করা হয়েছে যে, q এর সংখ্যা পূর্ব থেকে নির্ধারিত করা সহজ নয়। কাজেই $q=0$ বা 1 থেকে শুরু করে প্রতি ক্ষেত্রে পর্যাপ্ততা যাচাই করে পর্যায়ক্রমে q-এর সংখ্যা বাড়াতে হবে, যতক্ষণ না নাস্তিকল্পনা বাতিল হয়ে যায়। অবশ্য এরূপ পর্যায়ক্রমিক q-এর সংখ্যা বাড়াতে গিয়ে পর্যাপ্ততা যাচাই-এর জন্য যে কতগুলো নাস্তিকল্পনা যাচাই করা হয় তাদের জন্য বর্জনীয় মানের সমন্বয় করা হয় না বলে এই পদ্ধতি উত্তম নয়।

৫.৮ উপাদানের তাৎপর্য নির্ণয় (Interpretation of Factors)

উপাদান ভর নির্ণয় করার পর i-তম চলককে উপাদানের রৈখিক ফাংশন হিসেবে লেখা যায়

$$X_i = \sum_{j=1}^{q} \lambda_{ij} f_j \qquad (৫.৮.১)$$

এখানে X_i হলো আদর্শায়িত চলক এবং λ_{ij} সমূহ হলো উপাদান ম্যাট্রিক্স-এর i-তম সারির j-তম মান। এগুলোকে বলা হয় j-তম উপাদানের ভর। এই ভর ছোট হতে পারে, বড় হতে পারে, আবার শূন্যও হতে পারে। ভরগুলোর কোনো একটি যদি খুব বড় হয়, বা সব ভর বড় এবং ধনাত্মক হয়, বা ভরগুলো শূন্যের কাছাকাছি হয় এবং কিছু ভর শূন্য-এর কিছু ছোট বা কিছু বড় হয়, তাহলে উপাদানের তাৎপর্য নির্ণয় সহজতর হয়।

উপাদানসমূহ অসংশ্লেষিত হলে, উপাদান ভরগুলো চলক ও উপাদানসমূহের সংশ্লেষাঙ্ক নির্দেশ করে। কাজেই কোনো উপাদান ভর বড় হলে চলক ও ঐ উপাদানের সম্পর্ক বেশি বলে সিদ্ধান্ত নেয়া যায়। অর্থাৎ, λ_{ij} হলো i-তম চলক ও j-তম উপাদানের সংশ্লেষাঙ্ক। আগেই উল্লেখ করা হয়েছে যে $\hat{\Lambda} = [\lambda_{ij}]$ হলো উপাদান ধরন ম্যাট্রিক্স (factor pattern matrix)। উপাদানসমূহ অসংশ্লেষিত হলে উপাদান ধরন ম্যাট্রিক্স এবং উপাদান গঠন ম্যাট্রিক্স (factor structure matrix) একই হয়, এখানে উপাদান গঠন ম্যাট্রিক্স হলো চলকসমূহ ও উপাদানসমূহের মধ্যকার সংশ্লেষাঙ্কসমূহের ম্যাট্রিক্স। কাজেই λ_{ij} এর মানের ভিত্তিতে উপাদানসমূহকে গুচ্ছায়ন করা যেতে পারে এবং λ_{ij} এর মান বড় হলে j-তম উপাদানকে i-তম চলকসমূহের একটি সাধারণ চলক হিসেবে চিহ্নিত করা যেতে পারে।

যে কোনো উপাদানকে চলকসমূহের একটি সাধারণ চলক হিসেবে চিহ্নিত করার আগে চলকসমূহের ভেদের কত অংশ উপাদানসমূহ ব্যাখ্যা করতে পারে তা লক্ষ্য করা প্রয়োজন। সাধারণত j-তম উপাদান i-তম চলকের ভেদের $(\lambda_{ij}^2 \times 100)$ শতাংশ ব্যাখ্যা করে এবং উপাদানসমূহ i-তম চলকের ভেদের মোট $(\Sigma \lambda_{ij}^2 \times 100)$

শতাংশ ব্যাখ্যা করে। এটিই হলো i-তম চলকের কম্যুনালিটি। কম্যুনালিটির মান যতো বড় হবে উপাদান প্রতিকৃতি চলকসমূহকে ততো ভালভাবে ব্যাখ্যা করবে।

উপাদান ম্যাট্রিক্স-এর আরো একটি ব্যাখ্যা করা যায়। উপাদানসমূহ সংশ্লেষিত বা অসংশ্লেষিত যাই হোক না কেন উপাদান ভরগুলো হলো উপাদানসমূহের উপর i-তম চলকের আদর্শায়িত নির্ভরণ সহগ (regression coefficients)। উপাদানসমূহ অসংশ্লেষিত হলে λ_{ij} গুলো অনপেক্ষ হবে। সেক্ষেত্রে λ_{ij} গুলো i-তম চলকের ভেদে j-তম উপাদানের পরিমাণ নির্ণয় করতে সাহায্য করে।

উপাদান বিশ্লেষণ করার জন্য অনুমান করা যায় যে, চলকসমূহের মধ্যে কতগুলো সাধারণ উপাদানের (common factor) প্রভাব আছে বলেই ঐগুলো সংশ্লেষিত। কাজেই i-তম চলক ও j-তম উপাদানের সংশ্লেষাঙ্ককে i-তম ও k-তম ($i \neq k = 1, 2, \ldots, p$) চলকের সংশ্লেষাঙ্ক-এর নিরূপক পাওয়ার জন্য ব্যবহার করা যায়। উপাদানসমূহ অসংশ্লেষিত হলে i-তম ও k-তম চলকের সংশ্লেষাঙ্কের নিরূপক হবে

$$\hat{r}_{ik} = \sum_{j=1}^{q} r_{ij}\, r_{kj} \qquad (৫.৮.২)$$

এখানে $\hat{r}_{ij}$ হলো i-তম চলক ও j-তম উপাদানের সংশ্লেষাঙ্ক। কাজেই

$$\hat{r}_{ik} = \sum_{j=1}^{q} \lambda_{ij}\, \lambda_{kj}$$

এই $\hat{r}_{ik}$ হলো উপাদান প্রতিকৃতি ব্যবহার করে j-তম ও k-তম চলকের সংশ্লেষাঙ্ক। উক্ত সংশ্লেষাঙ্ক নমুনা হতেও পাওয়া যায়। ধরা যাক তা r_{ik}। তাহলে ($\hat{r}_{ik} - r_{ik}$) হলো অবশিষ্ট। এই অবশিষ্ট যতো ছোট হবে উপাদান প্রতিকৃতি উপাত্তের জন্য ততো পর্যাপ্ত হবে।

আগেই উল্লেখ করা হয়েছে যে λ_{ij} হলো i-তম চলকের j-তম উপাদানের সংশ্লেষাঙ্ক। তাছাড়া ($\lambda_{ij}^2 \times 100$) হলো j-তম উপাদান দ্বারা i-তম চলকের ভেদের ব্যাখ্যা করা অংশ। সুতরাং, উপাদান ম্যাট্রিক্স হতে j-তম উপাদান দ্বারা সকল চলকের ভেদের কি পরিমাণ ব্যাখ্যা করা যায় তা নির্ণয় করা যায়। অর্থাৎ j-তম উপাদানের মোট ভেদাঙ্ক হলো

$$\lambda_{1j}^2 + \lambda_{2j}^2 + \cdots \lambda_{pj}^2$$

এটি λ_j এর সমান।

৫.৯ উপাদান সাফলাঙ্ক (Factor Score)

উপাদান বিশ্লেষণের জন্য যে প্রতিকৃতি [প্রতিকৃতি (৫.২.১)] বিবেচনা করা হয় তা থেকেই বলা যায় যে চলকসমূহ কতগুলো অজানা উপাদানের ফাংশন। এর উল্টোটাও উপাদান বিশ্লেষণে সম্ভব। অর্থাৎ উপাদানসমূহকেও চলকসমূহের ফাংশন হিসেবে প্রকাশ করা যায়। এরূপ একটি ফাংশন হলো

$$f_j = W_{1j}X_1 + W_{2j}X_2 + \cdots + W_{pj}X_p \qquad (৫.৯.১)$$

এই প্রতিকৃতি থেকে যে কোনো নমুনা বিন্দুর p চলকের মানের ভিত্তিতে ঐ নমুনা বিন্দুর প্রাসঙ্গিক j-তম উপাদানের মান নিরূপণ করা যায়। এই নিরূপিত মানই হলো উপাদান সাফলাঙ্ক (factor score)।

ধরা যাক j-তম উপাদানের h-তম নিরূপিত মান হলো $\hat{f}_{hj}$ এবং এটি চলকসমূহের রৈখিক ফাংশন। অর্থাৎ

$$\hat{f}_{hj} = B_1X_{h1} + B_2X_{h2} + \ldots + B_pX_{hp} \qquad (৫.৯.২)$$

এখানে $B_1, B_2, \cdots, B_p$ নির্ভরণ সহগের ন্যায় এবং $X_{h1}, X_{h2}, \ldots, X_{hp}$ হলো h-তম নমুনা বিন্দু হতে প্রাপ্ত p চলকের মান সমীকরণ ৫.৯.২-কে লেখা যায়

$$\hat{F} = XB \qquad (৫.৯.৩)$$

এখানে $\hat{F}$ হলো $(n \times q)$ ম্যাট্রিক্স যার মানগুলো হলো q উপাদানের n মান, X হলো $(n \times p)$ উপাত্ত ম্যাট্রিক্স, B হলো q স্তম্ভবিশিষ্ট p নির্ভরাঙ্ক-এর ম্যাট্রিক্স। এই ম্যাট্রিক্স-এর j-তম স্তম্ভ হলো k-তম উপাদানের n মান নিরূপণ করার জন্য p নির্ভরাঙ্ক। এ পর্যায়ে X-এর পরিবর্তে আদর্শায়িত ম্যাট্রিক্স Z ব্যবহার করা হলে ৫.৯.৩-এর আকার হয়

$$\hat{F} = Z\beta \qquad (৫.৯.৪)$$

এখানে β হলো আদর্শায়িত নির্ভরাঙ্ক সহগ ম্যাট্রিক্স। এখন ৫.৯.৪-কে Z' দ্বারা প্রাক গুণ করে এবং উভয় পাশে n দ্বারা ভাগ করে পাওয়া যায়

$$\frac{1}{n}Z'\hat{F} = \frac{1}{n}Z'Z\beta$$

$$= R\beta$$

এখন R হলো নমুনা সংশ্লেষাঙ্ক ম্যাট্রিক্স এবং $\frac{1}{n}Z'\hat{F}$ হলো উপাদানসমূহ ও চলকসমূহের সংশ্লেষাঙ্ক ম্যাট্রিক্স। সুতরাং লেখা যায়

$$\Lambda = R\Phi$$

$$\rightarrow \Phi = R^{-1}\Lambda$$

$$\therefore \hat{F} = Z R^{-1} \Lambda \qquad (৫.৯.৫)$$

উদাহরণ ৫.১ ঃ নিচে Gaffar (1996)-এর কাজের ভিত্তিতে শিশু মৃত্যুকে প্রভাবিত করে এমন কতগুলো চলকের কিছু তথ্য দেয়া হলো।

ক্রমিক সংখ্যা	A	B	C	D	E	F	G	H	I
1	45	0	0	56	2	1	11	0	3
2	43	4	1	55	2	1	10	0	2
3	35	1	0	40	3	1	4	0	2
4	38	2	2	44	3	1	7	0	2
5	45	4	1	50	4	0	6	0	3
6	30	3	1	35	4	2	4	0	3
7	37	2	0	40	1	1	8	0	2
8	48	0	2	49	1	1	6	0	1
9	24	2	3	26	2	1	1	0	2
10	40	0	0	49	1	3	4	0	2
11	43	1	0	44	2	1	8	0	2
12	38	0	0	50	1	1	6	0	1
13	40	0	0	50	1	1	7	0	3
14	40	4	1	50	4	0	7	0	2
15	42	0	0	41	1	2	8	0	2

16	49	0	0	51	4	0	9	0	2
17	40	0	0	49	1	1	9	0	2
18	32	3	3	33	4	2	4	0	3
19	42	2	0	48	4	1	11	0	2
20	34	3	2	37	3	1	7	0	3
21	48	2	0	50	1	2	7	0	3
22	43	1	0	52	3	0	13	0	3
23	38	2	0	48	2	1	8	0	2
24	34	2	2	37	4	0	5	0	2
25	47	0	0	48	3	0	9	0	2
26	34	2	1	42	4	1	4	0	2
27	41	4	1	45	3	1	6	0	2
28	46	0	0	52	3	2	11	0	3
29	37	0	0	45	0	2	8	1	1
30	45	1	0	57	4	0	9	1	2
31	43	1	0	46	2	1	13	1	2
32	43	1	0	45	2	1	14	1	3
33	45	2	0	56	4	0	8	1	3
34	48	0	0	53	0	2	9	1	3
35	37	4	2	42	4	1	10	1	3
36	43	3	1	52	3	1	10	1	3
37	37	0	0	52	1	1	9	1	2

38	42	0	0	52	3	2	7	1	2
39	45	0	0	50	1	1	10	1	3
40	26	3	2	30	3	1	3	1	1
41	42	3	2	45	4	2	7	1	2
42	30	2	2	39	2	1	4	1	1
43	49	1	0	56	0	1	11	1	2
44	43	3	0	52	4	2	7	1	2
45	42	4	1	56	1	1	5	1	2
46	37	4	2	52	4	1	10	1	3
47	40	0	0	47	1	2	8	1	2
48	40	0	0	47	2	1	13	1	2
49	45	0	0	75	1	1	7	1	2
50	45	0	0	55	0	2	9	1	2
51	35	1	0	35	1	2	15	3	3
52	48	1	0	56	3	1	19	2	3
53	49	0	0	63	0	2	12	3	2
54	47	1	0	52	3	1	14	5	2
55	49	0	2	55	1	2	11	2	2
56	49	0	0	51	1	1	15	2	2
57	31	3	2	35	3	2	5	1	1

এখানে, A = মায়ের বয়স, B = মায়ের শিক্ষা, C = মায়ের পেশা
D = বাবার বয়স, E = বাবার শিক্ষা, F = বাবার পেশা

G = জীবিত জন্মগ্রহণ করা সন্তানের সংখ্যা

H = মৃত সন্তানের সংখ্যা, I = অর্থনৈতিক অবস্থা

Gaffar দেখিয়েছেন যে তাঁর বিশ্লেষিত নমুনা উপাত্ত হতে অর্থনৈতিক অবস্থা এবং মৃত সন্তানের সংখ্যা এর মধ্যে কোনো তাৎপর্যপূর্ণ সম্পর্ক আছে বলা যায় না। কাজেই অর্থনৈতিক অবস্থা বাদ দিয়ে কোনো চলকগুচ্ছ শিশু মৃত্যুর জন্য বেশি দায়ী তা নির্ধারণ করার জন্য উপাদান বিশ্লেষণ করা যেতে পারে। এমন হতে পারে যে {A, B, C, G} গুচ্ছ এবং {D, E, F, G} গুচ্ছ মৃত সন্তানের ভেদ পর্যালোচনা করার জন্য পর্যাপ্ত।

উপাদান বিশ্লেষণ করার শুরুতেই চলকসমূহের মধ্যে তাৎপর্যপূর্ণ সংশ্লেষণ আছে কিনা তা যাচাই করা দরকার। কারণ চলকসমূহ অসংশ্লেষিত হলে তাদের মধ্যে কোনো সাধারণ উপাদান বিদ্যমান আছে তা আশা করা যায় না। এখানে নাস্তিকল্পনা হলো

$$H_0 : P = I \qquad [P = \text{গণসমষ্টি সংশ্লেষাঙ্ক ম্যাট্রিক্স}]$$

এবং নমুনা সংশ্লেষাঙ্ক ম্যাট্রিক্স হলো

$$R = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \end{array} \begin{bmatrix} A & B & C & D & E & F & G \\ 1.0000 & & & & & & \\ -0.4023^{**} & 1.0000 & & & & & \\ -0.5569^{**} & 0.5835^{**} & 1.0000 & & & & \\ 0.5100 & -0.0066 & -0.1940 & 1.0000 & & & \\ -0.2386 & 0.5888^{**} & 0.3721^{**} & 0.1049 & 1.0000 & & \\ -0.0383 & -0.2043 & 0.0285 & -0.0774 & -0.3949^{*} & 1.0000 & \\ 0.5993^{**} & -0.3257^{**} & -0.4934^{**} & 0.2895 & -0.1533 & -0.0579 & 1.0000 \end{bmatrix}$$

$$* \Rightarrow P-\text{value} < 0.01\,, \quad ** \Rightarrow P-\text{value} < 0.001$$

$|R| = 0.0946704$। সুতরাং Box (1949) এর প্রস্তাব অনুযায়ী ৫.৫.৩ তথ্যজমানের ভিত্তিতে $-2\log\lambda = 124.55$ পাওয়া যায়। এখানে $P=7$, $n=57$। সুতরাং $-2\log\lambda$ এর স্বাধীনতার মাত্রা হলো 21 এবং $P(-2\log\lambda \geq 124.55) = 0.00000$ হওয়ায় চলকসমূহ সংশ্লেষিত বলে সিদ্ধান্ত গ্রহণ করা

যায়। এখানে অনুমান করা হচ্ছে যে উপাত্তসমূহ বহুচলক পরিমিত বিন্যাস হতে চয়ন করা হয়েছে।

আলোচিত উপাত্তের ক্ষেত্রে KMO = 0.72। সুতরাং এক্ষেত্রে উপাদান বিশ্লেষণ করা হলে তা মোটামুটি গ্রহণযোগ্য হবে। আবার বিশ্লেষণে অন্তর্ভুক্ত চলকসমূহ পর্যাপ্তও। কারণ A, B, C, D, E, F, G এর নমুনা পর্যাপ্ততার পরিমাপ হলো যথাক্রমে 0.71, 0.74, 0.79, 0.59, 0.68, 0.55 এবং 0.80। সবগুলো মানই বেশ বড়। কাজেই সব চলকই বিশ্লেষণে অন্তর্ভুক্ত হতে পারে।

যে কোনো একটি চলকের সাথে বাকি চলকসমূহের বহু-সংশ্লেষাঙ্ক এর বর্গকে (R^2) চলকসমূহের সংশ্লেষণের মাত্রা নির্ণয়ের জন্য ব্যবহার করা যায়। উপাদান বিশ্লেষণের ক্ষেত্রে এই মানকে বলা হয় কম্যুনালিটি। আলোচিত চলকসমূহের ক্ষেত্রে কম্যুনালিটি হলো যথাক্রমে 0.77589, 0.72372, 0.66369, 0.52527, 0.75562, 0.45413 এবং 0.60524। কোনো মানই ছোট নয়। সুতরাং আলোচিত চলকসমূহ বিশ্লেষণের জন্য পর্যাপ্ত।

এখন আলোচিত চলকসমূহ ব্যবহার করে উপাদান বিশ্লেষণ শুরু করার জন্য প্রাথমিক তথ্যজমানসমূহ (initial statistics) সারণি ৫.১-এ উপস্থাপন করা হলো। এখানে 7 চলকের ভিত্তিতে 7 উপাদান পাওয়া যেতে পারে। এই 7

সারণি ৫.১ঃ প্রাথমিক তথ্যজমানসমূহ।

চলক	কম্যুনালিটি	উপাদান	আইগেন মান	ভেদের শতকরা হার	ভেদের ক্রমসঞ্চিত শতকরা হার
A	1.00	1	2.86	40.9	40.9
B	1.00	2	1.64	23.4	64.3
C	1.00	3	0.90	12.8	77.2
D	1.00	4	0.57	8.1	85.3
E	1.00	5	0.40	5.7	91.0
F	1.00	6	0.33	4.7	95.8
G	1.00	7	0.30	4.2	100.0

উপাদান চলকসমূহের ভেদের 100% ব্যাখ্যা করতে পারে। প্রধান উপাদান পদ্ধতির মাধ্যমে প্রাপ্ত উপাদান ম্যাট্রিক্স সারণি ৫.২-এ উপস্থাপন করা হলো।

সারণি ৫.২ ঃ প্রধান উপাদান পদ্ধতি হতে প্রাপ্ত সকল উপাদানের জন্য উপাদান ম্যাট্রিক্স ।

চলক	উপাদান						
	1	2	3	4	5	6	7
A	0.807	0.353	0.144	−0.011	−0.125	−0.027	0.432
B	−0.749	0.404	0.202	0.188	−0.101	−0.436	0.020
C	−0.814	−0.022	0.290	0.106	−0.375	0.310	0.067
D	0.381	0.617	0.563	−0.321	−0.005	0.044	−0.229
E	−0.568	0.658	−0.028	0.152	0.410	0.198	0.116
F	0.136	−0.660	0.656	0.241	0.235	−0.020	0.051
G	0.708	0.323	−0.068	0.582	−0.109	0.065	−0.188

কিন্তু উপাদান বিশ্লেষণের মূল উদ্দেশ্য হলো গণসমষ্টির যে বৈশিষ্ট্য পর্যালোচনা করার জন্য একক (individual) গুলোর যে সকল চলক পর্যালোচনা করা হয় সে সকল চলকের পরিবর্তে অল্প কিছু উপাদান দ্বারা ঐ চলকসমূহের ভেদের বৃহত্তর অংশ প্রকাশ করা যায় কিনা তা বিশ্লেষণ করা। এখানে উপাদান-১ ও উপাদান-২ এর ভেদাঙ্ক ১ এর চেয়ে বড়। এই দুটি উপাদান চলকের মোট ভেদের 64.3% প্রকাশ করে থাকে। কাজেই এক্ষেত্রে দুটি উপাদান নির্ধারণ করা যায়। সারণি ৫.৩-এ এই দুটি উপাদানের জন্য উপাদান ম্যাট্রিক্স, কম্যুনালিটি এবং একক উপাদানের (unique factor) ভেদাঙ্ক উপস্থাপন করা হলো।

সারণি ৫.৩ ঃ প্রধান উপাদান পদ্ধতিতে প্রাপ্ত বিশ্লেষণের ফলাফল।

চলক	উপাদান		h_i^2	ψ_i^2
	1	2		
A	0.807	0.353	0.78	0.22
B	−0.749	0.404	0.72	0.28
C	−0.814	−0.022	0.66	0.34
D	0.381	0.617	0.53	0.47
E	−0.568	0.658	0.76	0.24
F	0.136	−0.660	0.45	0.55
G	0.708	0.323	0.61	0.39
মোট ভেদাঙ্ক (%)	40.9	23.4	64.3	35.7
সাধারণ ভেদাঙ্ক (%)	63.6	36.4		
ভেদে j-তম উপাদানের পরিমাণ (λ)	2.86	1.64		

উপরে প্রাপ্ত উপাদান ভর (factor loading) ম্যাট্রিক্স হতে পাওয়া যায়

$$A = \quad 0.807\ f_1 + 0.353\ f_2$$

$$B = -0.749\ f_1 + 0.404\ f_2$$

$$C = -0.814\ f_1 - 0.022\ f_2$$

$$D = \quad 0.381\ f_1 + 0.617\ f_2$$

$$E = -0.568\ f_1 + 0.658\ f_2$$

$$F = \quad 0.136\ f_1 - 0.66\ f_2$$

$$G = \quad 0.708\ f_1 + 0.323\ f_2$$

দেখা যাচ্ছে যে, f_1 চলক A, B, C, E এবং G এর সাথে বেশি সংশ্লেষিত এবং f_2 চলক D, E, F এর সাথে বেশি সংশ্লেষিত। এখানে f_1 চলক A, B, C, E এবং G এর 65.1%, 56.1%, 66.3%, 32.2% এবং 50.1%, যথাক্রমে ভেদ ব্যাখ্যা করেছে। উপাদান f_1 এবং f_2 দ্বারা A চলকের 77.6% ভেদ প্রকাশিত হয়েছে বাকি 22.4% ভেদ প্রকাশিত হয়নি। এই 22.4% হলো চলকের এককতার (uniqueness) পরিমাণ। উপাদান f_1 চলক A, B, C, E এবং G এর প্রতিটির কমপক্ষে 25% (Dillon and Goldstein) ভেদ ব্যাখ্যা করে বলে এই চলকগুচ্ছ এর পরিবর্তে উপাদান-১ এবং D, E, F গুচ্ছের পরিবর্তে উপাদান-২ ব্যবহার করে শিশু মৃত্যুর ভেদ পর্যালোচনা করা যেতে পারে।

আগেই আলোচনা করা হয়েছে যে, উপাদানগুলো অনপেক্ষ হলে i-তম ও k-তম চলকের মধ্যে সংশ্লেষাঙ্ক $\hat{r}_{ik}$ এর মান পাওয়ার জন্য i-তম চলক ও j-তম উপাদানের উপাদান ভর ব্যবহার করা যায়। আলোচিত উপাত্তের ক্ষেত্রে ৫.৮.২ সূত্র প্রয়োগ করে চলক A ও B এর সংশ্লেষাঙ্ক-এর নিরূপক হবে

$$\hat{r}_{12} = 0.807 \times (-0.749) + 0.353 \times 0.404 = -0.462$$

অপরপক্ষে নমুনা হতে প্রাপ্ত $r_{12} = -0.402$। এখানে অবশিষ্ট হলো $r_{12} - \hat{r}_{12}$ $= 0.06$। অনুরূপভাবে সকল চলকের জন্য জোড়ায় জোড়ায় সংশ্লেষাঙ্ক $[\hat{r}_{ik}]$ এবং ঐগুলোর প্রাসঙ্গিক অবশিষ্টসমূহ নির্ণয় করা যায়। সারণি ৫.৪-এ এই মানগুলো উপস্থাপন করা হলো। এই সারণির কৌণের (diagonal) উপরের মানগুলো হলো অবশিষ্ট এবং কৌণের নিচের মানগুলো হলো নিরূপিত সংশ্লেষাঙ্ক এবং কৌণের মানগুলো হলো কম্যুনালিটি।

সারণি ৫.৪ ঃ নিরূপিত সংশ্লেষাঙ্ক এবং অবশিষ্ট।

	A	B	C	D	E	F	G
A	0.776	0.060	0.108	−0.015	−0.012	0.085	−0.086
B	−0.462	0.724	−0.018	0.030	−0.102	0.164	0.074
C	−0.665	0.601	0.664	0.129	−0.076	0.125	0.090
D	0.525	−0.036	−0.323	0.525	−0.084	0.278	−0.179
E	−0.227	0.691	0.449	0.189	0.756	0.116	0.037
F	−0.123	−0.368	−0.096	−0.355	−0.511	0.454	0.059
G	0.685	−0.400	−0.584	0.469	−0.190	−0.117	0.605

লক্ষ্য করা যাচ্ছে যে অধিকাংশ ক্ষেত্রে অবশিষ্ট-এর মান বড় (75.2% ক্ষেত্রে চিহ্ন-বর্জিত অবশিষ্ট 0.05 এর বড়)। এক্ষেত্রে দুই উপাদানবিশিষ্ট উপাদান মডেল আলোচ্য নমুনার জন্য পর্যাপ্ত বলা যায় না।

সারণি ৫.১ থেকে লক্ষ্য করা যাচ্ছে যে, প্রথম তিনটি উপাদান চলকসমূহের ভেদের 77.2% ব্যাখ্যা করে থাকে এবং তৃতীয় উপাদানটির ভেদাঙ্ক প্রায় 1। সুতরাং বিকল্প হিসেবে তিন উপাদানবিশিষ্ট মডেল বিবেচনা করা হলে নিরূপিত সংশ্লেষাঙ্ক এবং অবশিষ্ট হবে নিম্নরূপ : এই প্রতিকৃতির ক্ষেত্রে সংশ্লেষাঙ্ক-এর

সারণি ৫.৫ ঃ তিন উপাদানবিশিষ্ট মডেলের জন্য সংশ্লেষাঙ্ক এবং অবশিষ্ট।

	A	B	C	D	E	F	G
A	0.797	0.031	0.066	−0.096	−0.008	−0.009	−0.076
B	−0.433	0.764	−0.077	−0.084	−0.096	0.032	0.088
C	−0.623	0.660	0.748	−0.034	−0.068	−0.066	0.110
D	0.606	0.077	−0.160	0.842	−0.069	−0.091	−0.141
E	−0.231	0.685	0.440	0.174	0.756	0.135	0.035
F	−0.029	−0.236	0.094	0.014	−0.530	0.884	0.104
G	0.675	−0.414	−0.603	0.430	−0.188	−0.162	0.610

অবশিষ্টের পরিমাণ কিছুটা কমে থাকলেও তেমন কোনো উল্লেখযোগ্য উন্নতি পরিলক্ষিত হয়নি। তবু, যেহেতু উপাদানসমূহ চলকের বেশি ভেদ ব্যাখ্যা করতে পারে, সে কারণে তিন উপাদানবিশিষ্ট বিবেচনা প্রতিকৃতি করা হলে অযৌক্তিক হবে না।

৫.১০ উপাদানের রোটেশন (Rotation of factors)

উপাদানসমূহের তাৎপর্য আলোচনা করতে গিয়ে উল্লেখ করা হয়েছে যে, λ_{ij} হলো j-তম উপাদান ও i-তম চলকের সংশ্লেষাঙ্ক। এই সংশ্লেষাঙ্ক যত বড় হবে j-তম উপাদান j-তম চলকের ততো বেশি ভেদ ব্যাখ্যা করতে পারে। যেহেতু যে কোনো উপাদান কতগুলো এক জাতীয় চলকের গুচ্ছের ভেদ বেশি পরিমাণে ব্যাখ্যা করতে পারে, সে কারণে λ_{ij} এর মান পর্যালোচনা গুরুত্বপূর্ণ। সাধারণত λ_{ij}^2 এর মান 0.25 বা তার উপর হলেই বিবেচনা করা হয় যে, j-তম উপাদান i-তম চলকের ভেদ ব্যাখ্যা করতে সহায়ক। কিন্তু অনেক ক্ষেত্রে λ_{ij}^2 এর মান 0.25 এর চেয়ে সামান্য ছোট হতে পারে। সেক্ষেত্রে সোজাসুজি সিদ্ধান্ত সহজ নয়। সে কারণে যে কোনো উপাদানকে কোনো একগুচ্ছ চলকের ভেদ ব্যাখ্যা করার সহায়ক বিবেচনা করার জন্য উপাদানসমূহের রোটেশন করা দরকার। এই রোটেশন উপাদান ম্যাট্রিক্স-এর নতুন মান পেতে সাহায্য করে এবং ভিন্ন ভিন্ন উপাদান ভিন্ন ভিন্ন চলকগুচ্ছের ভেদ ব্যাখ্যা করার জন্য চিহ্নিত করার সহায়ক হয়।

ধরা যাক একটি কাল্পনিক উপাদান ম্যাট্রিক্স হলো নিম্নরূপ : এখানে λ_{ij} এর প্রতিটি মানই বড় এবং উভয় উপাদানই সকল চলকের ভেদ ব্যাখ্যা করার সহায়ক। এখানে উপাদান বিশ্লেষণ হতে উপাত্ত সংকোচন করার সুযোগ নেই। অপরপক্ষে যদি উপাদান ম্যাট্রিক্সটি নিম্নরূপ হয়, তাহলে এটি পরিস্কার যে উপাদান-1 চলক-1 ও

কাল্পনিক উপাদান ম্যাট্রিক্স

চলক	উপাদান	
	1	2
1	0.489	0.503
2	0.611	−0.689
3	0.700	0.600
4	0.653	0.700

চলক-2 এর বেশি ভেদ ব্যাখ্যা করতে সহায়ক এবং উপাদান-2 চলক-3 ও চলক-4 এর বেশি ভেদ ব্যাখ্যা করতে সহায়ক। এরূপ একটি সহজ উপাদান ম্যাট্রিক্স পাওয়ার জন্যই উপাদান রোটেশন করা প্রয়োজন।

চলক	উপাদান	
	1	2
1	0.725	0.032
2	0.674	0.124
3	0.055	0.801
4	0.041	0.635

সমকৌণিক রোটেশনের জন্য অনেক পদ্ধতি আছে। এগুলোর মধ্যে উল্লেখযোগ্য পদ্ধতি হলো (১) Varimax পদ্ধতি, (২) Quartimax পদ্ধতি, (৩) Equamax পদ্ধতি। এই পদ্ধতিগুলো প্রয়োগ করা হলে উপাদানসমূহের সমকৌণিকতা (orthogonality) নষ্ট হয় না। অনেক সময় উপাদানের সমকৌণিকতার বিষয় বিবেচনা না করেও রোটেশন করা হয়। এরূপ একটি রোটেশনের নাম হলো Oblique rotation।

Varimax Rotation : Kaiser (1958) এ রোটেশনের প্রস্তাব করেছেন। এ রোটেশনের সাহায্যে অল্প কিছু উপাদান ভর খুব বড় পাওয়া যায় এবং বেশির ভাগ উপাদান ভর শূন্য বা এর কাছাকাছি হয়। এ সম্পর্কে আরো বিস্তারিত জানার জন্য Harman (1976) এবং Rummel (1970) পর্যালোচনা করা যেতে পারে।

Varimnx পদ্ধতি সাধারণত প্রধান উপাদান পদ্ধতিতে উপাদান বিশ্লেষণ করা হলে ঐ উপাদান ভর রোটেট করার জন্য প্রয়োগ করা হয়।

ধরা যাক Λ হলো $(p \times q)$ উপাদান ভর ম্যাট্রিক্স এবং G হলো $(q \times q)$ সমকৌণিক ম্যাট্রিক্স। যেহেতু Varimax পদ্ধতিতে উপাদানসমূহ সমকৌণিক হয়, সে কারণে Λ-এর উপর একটি সমকৌণিক পরিবর্তন করা যাক। ধরা যাক পরিবর্তিত উপাদান ভর ম্যাট্রিক্স হলো Δ, যেখানে

$$\Delta = \Lambda G = (\delta_{ij}) \qquad (৫.১০.১)$$

এবং δ_{ij} হলো i-তম চলকের জন্য j-তম উপাদানের ভর।

ধরা যাক δ_{ij} কে i-তম চলকের কম্যুনালিটি দ্বারা নরমালাইজড্ করা হয়েছে। অর্থাৎ δ_{ij}-কে পরিবর্তন করে $d_{ij} = \delta_{ij}/h_i$ পাওয়া গেছে। এখন j-তম উপাদানের নরমালাইজড্ উপাদানে ভরের বর্গের ভেদাঙ্ক পাওয়া যায়

$$\sum_{i=1}^{p}(d_{ij}^2-\bar{d}_j)^2 \quad \text{এখানে } \bar{d}_j=\sum_{i=1}^{p} d_{ij}^2/p$$

সকল চলকভিত্তিক এই ভেদাঙ্কের যোগফল হলো

$$\varphi=\sum_{i=1}^{p}\sum_{j=1}^{q}(d_{ij}^2-\bar{d}_j)^2$$

Varimax পদ্ধতির বৈশিষ্ট্য হলো এটি φ-কে সর্বোত্তম করে এবং φ হলো G-এর ফাংশন। Kaiser সমকৌণিক ম্যাট্রিক্স G নির্ণয়ের পদ্ধতি ব্যাখ্যা করেছেন যা φ-কে সর্বোত্তম করে। এই পদ্ধতি হলো একটি ইটারেটিভ পদ্ধতি।

Varimax পদ্ধতিতে মূলত উচ্চ ভরবিশিষ্ট চলকের সংখ্যা ন্যূনতম করে। কিন্তু এটি উপাদান প্রতিকৃতির যথাযথতা যাচাইকে প্রভাবিত করে না। এই পদ্ধতিতে কম্যুনালিটিসমূহের পরিবর্তন হয় না, উপাদান দ্বারা ব্যাখ্যা করা ভেদের পরিমাণে কোনো পরিবর্তন হয় না কিন্তু প্রতি উপাদান দ্বারা ব্যাখ্যা করা ভেদের পরিমাণে পরিবর্তন হয়ে থাকে। এই পদ্ধতিতে উপাদান সংশ্লেষাঙ্ক ম্যাট্রিক্স আইডেনটিটি ম্যাট্রিক্স হয়।

সমকৌণিক পদ্ধতিগুলোর মধ্যে অন্য একটি হলো Quartimax পদ্ধতি। এই পদ্ধতিতে উচ্চ ভরবিশিষ্ট চলকসমূহের ভর মধ্যমানের ভরে পরিবর্তিত হয়ে থাকে। আবার Equamax পদ্ধতি হলো Varimax এবং Quartimax পদ্ধতির মিশ্রণ।

Oblique Rotation : এই পদ্ধতিতে উচ্চ ভর এবং নিম্ন ভরগুলো বৃদ্ধি পায় এবং মধ্যমানের ভরগুলো কমতে থাকে। Rummel (1970) এই রোটেশনের জন্য ব্যবহৃত পদ্ধতিসমূহের সূত্রসহ সকল বিষয় আলোচনা করেছেন।

এই রোটেশনের ক্ষেত্রেও কম্যুনালিটি-এর পরিবর্তন হয় না। কিন্তু উপাদান ভরসমূহ উপাদান ও চলকসমূহের সংশ্লেষাঙ্ক নির্দেশ করে না। অবশ্য উপাদান ভরগুলো আংশিক নির্ভরাঙ্ক নির্দেশ করে। কাজেই এই পদ্ধতিতে উপাদান ভর ম্যাট্রিক্স এবং উপাদান গঠন ম্যাট্রিক্স ভিন্ন হয়ে থাকে এবং উপাদান সংশ্লেষাঙ্ক ম্যাট্রিক্স আইডেনটিটি ম্যাট্রিক্স হয় না।

Varimax রোটেশন অনুসারে উদাহরণ ৫.১-এর উপাত্তের জন্য রোটেটেড উপাদান ম্যাট্রিক্স, কম্যুনালিটি ও একক উপাদানের ভেদাঙ্ক সারণি ৫.৬-এ উপস্থাপন করা হলো। এই বিশ্লেষণ হতে একটি বিষয় পরিষ্কার যে, শিশু মৃত্যুর ভেদ পর্যালোচনার জন্য A, D এবং G মিলে একটি চলকগুচ্ছ হতে পারে এবং B, E ও F মিলে আরেকটি চলকগুচ্ছ হতে পারে। কারণ A, D এবং G f_1 এর সাথে

সারণি ৫.৬ ঃ Varimax রোটেশন পদ্ধতিতে প্রাপ্ত ফলাফল।

চলক	উপাদান			
	1	2	h_i^2	ψ_i^2
A	0.875	−0.098	0.78	0.22
B	−0.447	0.724	0.72	0.28
C	−0.716	0.388	0.66	0.34
D	0.638	0.344	0.53	0.47
E	−0.164	0.854	0.76	0.24
F	−0.212	−0.640	0.45	0.55
G	0.774	−0.074	0.61	0.39
মোট ভেদাঙ্ক (%)	40.9	23.4	64.3	35.7
সাধারণ ভেদাঙ্ক (%)	63.6	36.4		
ভেদে j-তম উপাদানের পরিমাণ (λ)	2.86	1.64		

ধনাত্মকভাবে সংশ্লেষিত এবং B ও E f_2 এর সাথে ধনাত্মকভাবে সংশ্লেষিত। এখানে মূলত (১) বয়স এবং (২) শিক্ষা শিশু মৃত্যুর ভেদ পর্যালোচনার জন্য যথেষ্ট। অনুরূপ সিদ্ধান্ত equamax রোটেশন থেকেও করা যায়।

আলোচিত উদাহরণের ক্ষেত্রে উপাদান সাফল্যাঙ্ক ম্যাট্রিক্স (Factor score coefficient matrix) সারণি ৫.৭-এ দেয়া হলো।

সারণি ৫.৭ ঃ উপাদান সাফল্যাঙ্ক ম্যাট্রিক্স।

চলক	উপাদান-1	উপাদান-2
A	0.352	0.046
B	−0.104	0.344
C	−0.253	0.131
D	0.303	0.259
E	0.028	0.447
F	−0.160	−0.372
G	0.312	0.047

ধরা যাক চলকসমূহের একটি আদর্শায়িত মান হলো

$A = 0.691$, $B = -0.999$, $C = -0.687$, $D = 0.794$, $E = -0.180$

$F = -0.269$, $G = 0.739$ । তাহলে, উপাদান-1 এর সাফল্যাঙ্ক হবে

$$0.352 \times 0.691 + 0.104 \times 0.999 + 0.253 \times 0.687 + 0.303 \times 0.794$$
$$-0.028 \times 0.180 + 0.16 \times 0.269 + 0.312 \times 0.739 = 1.030$$

অনুরূপভাবে, উপাদান-2 এর সাফল্যাঙ্ক হলো -0.142 ।

উপরে প্রাপ্ত উপাদান সাফল্যাঙ্ক ম্যাট্রিক্স R^{-1} ও A গুণ করে পাওয়া। এখানে

	A	B	C	D	E	F	G
$R^{-1} = A$	2.393						
B	0.292	2.073					
C	0.513	-0.767	2.015				
D	-0.908	-0.212	0.016	1.487			
E	0.228	-0.778	-0.220	-0.282	1.825		
F	0.111	0.137	-0.258	-0.075	0.549	1.252	
G	-0.782	0.072	0.384	0.005	-0.106	0.029	1.666

উদাহরণ ৫.১-এর ক্ষেত্রে oblique রোটেশন হতে প্রাপ্ত ফলাফল সারণি ৫.৮-এ উপস্থাপন করা হলো। লক্ষ্য করা যাচ্ছে যে, এই বিশ্লেষণের ক্ষেত্রে উপাদান

সারণি ৫.৮ ঃ Oblique রোটেশন হতে প্রাপ্ত উপাদান ধরন ম্যাট্রিক্স ও উপাদান গঠন ম্যাট্রিক্স ক ।

চলক	উপাদান ধরন ম্যাট্রিক্স		উপাদান গঠন ম্যাট্রিক্স	
	1	2	1	2
A	−0.876	−0.042	−0.880	−0.116
B	0.434	0.696	0.493	0.733
C	0.711	0.343	0.740	0.403

D	−0.647	0.385	−0.615	0.331
E	0.146	0.845	0.218	0.857
F	0.226	−0.654	0.171	−0.635
G	−0.776	−0.025	−0.778	−0.090

ভরগুলো (উপাদান ধরন ম্যাট্রিক্স) উপাদান ও চলকসমূহের সংশ্লেষাঙ্ক (উপাদান গঠন ম্যাট্রিক্স) নির্দেশ করে না। কিন্তু নিরূপিত কম্যুনালিটি এর কোনো পরিবর্তন হয় না। যেমন, চলক A এর প্রাসঙ্গিক কম্যুনালিটি হলো

$$(-0.88)^2+(-0.116)^2=0.78$$

উদাহরণ ৫.২ ঃ উদাহরণ ৫.১-এর উপাত্তের ক্ষেত্রে সর্বোত্তম সম্ভাব্য পদ্ধতি (ML method) প্রয়োগ করে উপাদান বিশ্লেষণ করা যাক।

সর্বোত্তম সম্ভাব্য পদ্ধতিতে উপাদান বিশ্লেষণ করার তাত্ত্বিক তথ্য ৫.৩ অনুচ্ছেদে আলোচনা করা হয়েছে। ব্যবহারিক ক্ষেত্রে প্রধান উপাদান পদ্ধতি ব্যতীত অন্য কোনো পদ্ধতি প্রয়োগ করা হলে উপাদান বিশ্লেষণ হতে ভিন্নতর ফল পাওয়া যায়। তবে প্রাথমিক তথ্যজমান (initial statistics) গুলো প্রায় সবই এক। শুধু কম্যুনালিটি এর পরিবর্তন হয়। প্রধান উপাদান পদ্ধতিতে প্রাথমিক কম্যুনালিটিগুলো 1 ধরা হয়। কিন্তু অন্য পদ্ধতিতে প্রাথমিক কম্যুনালিটি হলো R_i^2, এখানে R_i^2 হলো i-তম চলকের সাথে অন্যান্য চলকের বহু-সংশ্লেষাঙ্ক এর বর্গ। এখন ML

সারণি ৫.৯ ঃ সর্বোত্তম সম্ভাব্য পদ্ধতির ক্ষেত্রে প্রাথমিক তথ্যজমানসমূহ।

চলক	কম্যুনালিটি	উপাদান	আইগেন মান	ভেদের শতকরা হার	ভেদের ক্রম-যোজিত শতকরা হার
A	0.58219	1	2.86	40.9	40.9
B	0.51771	2	1.64	23.4	64.3
C	0.50379	3	0.90	12.8	77.2
D	0.32767	4	0.57	8.1	85.3
E	0.45203	5	0.40	5.7	91.0
F	0.20138	6	0.33	4.7	95.8
G	0.39960	7	0.30	4.2	100.0

পদ্ধতিতে উপাদান বিশ্লেষণ করার জন্য প্রাথমিক তথ্যজমানসমূহ সারণি ৫.৯-এ উপস্থাপন করা হলো। সারণি ৫.১০-এ উপাদান ম্যাট্রিক্সসহ বিস্তারিত ফলাফল উপস্থাপন করা হলো। দেখা যাচ্ছে যে, ML পদ্ধতি দুটি উপাদান নির্ধারণ

সারণি ৫.১০ ঃ সর্বোত্তম সম্ভাব্য পদ্ধতিতে বিশ্লেষিত ফলাফল।

চলক	উপাদান ম্যাট্রিক্স 1	2	কম্যুনালিটি	উপাদান	আইগেন মান	ভেদের শতকরা হার	ভেদের ক্রমযোজিত শতকরা হার
A	0.836	0.317	0.80	1	2.47	35.3	35.3
B	-0.675	0.483	0.69	2	1.13	16.1	51.4
C	-0.728	0.104	0.54				
D	0.383	0.503	0.40				
E	-0.476	0.582	0.57				
F	0.074	-0.388	0.16				
G	0.638	0.206	0.50				

করেছে। কিন্তু উপাদান দুটি মোট ভেদের মাত্র 51.4% ব্যাখ্যা করেছে। প্রথম উপাদান ব্যাখ্যা করেছে 35.3% এবং দ্বিতীয় উপাদান ব্যাখ্যা করেছে 16.1%। এখানে প্রাথমিক তথ্যজমান হতে লক্ষ্য করা যাচ্ছে যে উপাদান-1 40.9% এবং উপাদান-2 23.4% ভেদ ব্যাখ্যা করতে পারে। কিন্তু ML পদ্ধতি প্রয়োগ করার পর ভেদের এই শতকরা হারে পরিবর্তন লক্ষ্য করা যাচ্ছে। এরূপ পরিবর্তন অবশ্য অন্য কোনো পদ্ধতিতে হয় না।

ML পদ্ধতিতে প্রাপ্ত দুটি উপাদান প্রতিকৃতির জন্য যথেষ্ট। এটি χ^2 যাচাই তথ্যজমান ৫.৭.১-এর মাধ্যমে বুঝা যায়। এখানে $p(\chi^2 \geq 8.52)=0.3844$। এর স্বাধীনতার মাত্রা হলো 8।

ষষ্ঠ অধ্যায়

কানুনী সংশ্লেষণ বিশ্লেষণ
(Canonical Correlation Analysis)

৬.১ সূচনা

যে কোনো গবেষণায় কোনো নমুনা বিন্দু হতে একাধিক চলকের মান নথিভুক্ত করা হয় এবং একাধিক নমুনা বিন্দু হতে নথিভুক্ত তথ্যের ভিত্তিতে চলকসমূহের পারস্পরিক সম্পর্ক পর্যালোচনা গবেষণার একটি গুরুত্বপূর্ণ বিষয়। দুই-এর অধিক চলকের পারস্পরিক সম্পর্ক বহু-নির্ভরণ বিশ্লেষণের মাধ্যমে করা হয়। সেক্ষেত্রে একটি নির্ভরশীল চলকের ভেদ ব্যাখ্যা করার জন্য একাধিক অনির্ভরশীল চলক ব্যবহার করা হয়। বাস্তব ক্ষেত্রে একাধিক নির্ভরশীল চলক একাধিক অনির্ভরশীল চলকের দ্বারা প্রভাবিত হতে পারে। যেমন, Pulmonate slugs এর ক্ষেত্রে তাদের শারীরিক ওজন এবং শেলের বিস্তার শারীরিক দৈর্ঘ্য এবং শেলের দৈর্ঘ্যের উপর নির্ভরশীল। ব্যবসা ক্ষেত্রে একটি কারখানায় উৎপাদিত দ্রব্যসমূহ ঐ কারখানায় নিয়োজিত শ্রমিক, বিনিয়োগকৃত মূলধন, দ্রব্যসমূহের চাহিদা ইত্যাদি চলকসমূহের উপর নির্ভরশীল। এক্ষেত্রে একটি হলো নির্ভরশীল চলক গুচ্ছ যা সাধারণত Y-গুচ্ছ দ্বারা চিহ্নিত হয় এবং অন্যটি হলো অনির্ভরশীল চলক গুচ্ছ যা X-গুচ্ছ নামে পরিচিত। কানুনী সংশ্লেষণ বিশ্লেষণ হলো এমন একটি বহু-চলক পদ্ধতি যার মাধ্যমে Y-গুচ্ছ ও X-গুচ্ছ চলকের সংশ্লেষণ পর্যালোচনা করা হয়। উভয় গুচ্ছের চলকগুলো মানগত (Quantitative) এবং/বা গুণগত (Qualitative) চলক হতে পারে।

ধরা যাক Y-গুচ্ছ চলকের একটি রৈখিক সমাবেশ (Linear combination) হলো $Y^* = b'Y$ এবং X-গুচ্ছ চলকের একটি রৈখিক সমাবেশ হলো $X^* = a'X$। কানুনী সংশ্লেষণ বিশ্লেষণের উদ্দেশ্য হলো X^* ও Y^* এমনভাবে পেতে হবে যেন রৈখিক সমাবেশ দুটির সংশ্লেষণ বৃহত্তম হয়। সেদিক থেকে বলা যায় যে, কানুনী সংশ্লেষণ বিশ্লেষণ হলো বহু-চলক নির্ভরণ বিশ্লেষণের একটি বর্ধিত রূপ। পার্থক্য হলো যে, বহু-চলক নির্ভরণ বিশ্লেষণে অনির্ভরশীল চলক গুচ্ছের এমন একটি রৈখিক সমাবেশ নির্ণয় করা হয় যা একটি নির্ভরশীল চলকের ভেদের বৃহত্তর অংশ ব্যাখ্যা করতে পারে। ধরা যাক X-গুচ্ছে q চলক আছে এবং Y-গুচ্ছে p চলক আছে। বহু-নির্ভরণ বিশ্লেষণের ক্ষেত্রে $p = 1$। কিন্তু কানুনী সংশ্লেষণ বিশ্লেষণের ক্ষেত্রে $p \geq 1$।

এই বিশ্লেষণের সাথে প্রধান উপাদান বিশ্লেষণ (Principal component analysis) এর ধর্মের কিছুটা মিল রয়েছে। শেষোক্ত বিশ্লেষণের ক্ষেত্রে একগুচ্ছ চলকের রৈখিক সমাবেশ নির্ণয় করা হয় যে সমাবেশ আদি চলকসমূহের ভেদের বৃহত্তর অংশ ব্যাখ্যা করতে পারে। কিন্তু কানুনী সংশ্লেষণের ক্ষেত্রে দুটি চলক গুচ্ছের রৈখিক সমাবেশ নির্ণয় করা হয় এমনভাবে যেন সমাবেশদ্বয় সবচেয়ে বেশি সংশ্লেষিত হয়।

৬.২ গণসমষ্টি উপাত্ত হতে কানুনী সংশ্লেষণ বিশ্লেষণ (Canonical Correlation Analysis from Population Data)

ধরা যাক X-গুচ্ছে q চলক আছে এবং Y-গুচ্ছে p চলক আছে। অর্থাৎ

$$X' = [X_1, X_2 \cdots X_q] \text{ এবং } Y' = [Y_1, Y_2 \cdots Y_p]$$

এখানে X ও Y হলো যথাক্রমে q মাত্রার ও p মাত্রার ভেক্টর। আরো ধরা যাক যে X ও Y ভেক্টরের গড় ভেক্টর হলো যথাক্রমে μ_x ও μ_y। তাহলে তাদের ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্সসমূহকে লেখা যায়

$$\Sigma_{XX} = E[(X-\mu_x)(X-\mu_x)']$$

$$\Sigma_{YY} = E[(Y-\mu_y)(Y-\mu_y)']$$

$$\Sigma_{YY} = E[(X-\mu_x)(Y-\mu_y)']$$

ধরা যাক $Z=(X\ Y)$ হলো $(q+p)$ মাত্রার চলক। যেখানে

$$\Sigma_{ZZ} = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}$$

এখানে Σ_{XX}, Σ_{XY}, Σ_{YX} এবং Σ_{YY} এর অর্ডার হলো যথাক্রমে $(q \times q)$ $(q \times p)$, $(p \times q)$ এবং $(p \times p)$।

ধরা যাক X এর রৈখিক সমাবেশ হলো

$$X^* = a'X = a_1X_1 + a_2X_2 + \cdots + a_qX_q \qquad (৬.২.১)$$

এবং Y এর রৈখিক সমাবেশ হলো

$$Y^* = b'Y = b_1Y_1 + b_2Y_2 + \cdots + b_pY_p \qquad (৬.২.২)$$

কানুনী সংশ্লেষণ বিশ্লেষণের উদ্দেশ্য হলো $a'X$ ও $b'Y$ এমনভাবে নির্ণয় করতে হবে যেন তাদের সংশ্লেষণ বৃহত্তম হয়। ধরা যাক $a'X$ ও $b'Y$ এর সংশ্লেষাঙ্ক হলো

$$\rho(a, b) = \frac{a'\Sigma_{XY}b}{[(a'\Sigma_{XX}a)(b'\Sigma_{YY}b)]^{1/2}} \qquad (৬.২.৩)$$

এখানে a এবং b হলো কানুনী ভর (canonical weights) এর ভেক্টর। এই ভেক্টরগুলোর মানগুলো কানুনী চলক পেতে আদি চলকসমূহের গুরুত্ব নির্দেশ করে। এখানে $a'X$ ও $b'Y$ এর মান এমনভাবে নির্ণয় করতে হবে যেন $\rho(a, b)$ সর্বোত্তম হয়। জানা আছে যে সংশ্লেষাঙ্ক মাপনী বা আদি বিন্দু দ্বারা পরিবর্তিত হয় না। সুতরাং a ও b এর মাপনী (scale) পরিবর্তন করা হলে $\rho(a, b)$ এর মানে কোনো পরিবর্তন হবে না। সে কারণে a ও b এর একটি যে কোনো পরিবর্তন করা যেতে পারে যেন

$$a'\Sigma_{XX}a = b'\Sigma_{YY}b = 1 \text{ এবং } E(a'X) = E(b'Y) = 0$$

হয়। এখানে a ও b এর উপর উপরিউক্ত শর্ত আরোপ করা নিম্নলিখিত সমীকরণদ্বয়

$$\left(\sum_{XX}^{-1}\sum_{XY}\sum_{YY}^{-1}\sum_{YX} - \lambda I\right)a = 0 \qquad (৬.২.৪)$$

$$\left(\sum_{YY}^{-1}\sum_{YX}\sum_{XX}^{-1}\sum_{XY} - \lambda I\right)b = 0 \qquad (৬.২.৫)$$

সমাধান করার সমতুল। এখানে I হলো আইডেনটিটি ম্যাট্রিক্স, λ হলো নিয়ামক সমীকরণদ্বয় (characteristic equation)

$$\left|\sum_{XX}^{-1}\sum_{XY}\sum_{YY}^{-1}\sum_{YX} - \lambda I\right| = 0 \qquad (৬.২.৬)$$

$$\left|\sum_{YY}^{-1}\sum_{YX}\sum_{XX}^{-1}\sum_{XY} - \lambda I\right| = 0 \qquad (৬.২.৭)$$

এর বৃহত্তম আইগেন মান। এই বৃহত্তম আইগেন মানই হলো কানুনী সংশ্লেষাঙ্ক (Canonical correlation co-efficient) এর বর্গ। যেহেতু λ হলো

$$\sum_{XX}^{-1}\sum_{XY}\sum_{YY}^{-1}\sum_{YX} \quad \text{এবং} \quad \sum_{YY}^{-1}\sum_{YX}\sum_{XX}^{-1}\sum_{XY}$$

এর আইগেন মান, সে কারণে λ এর প্রাসঙ্গিক দুটি আইগেন ভেক্টর পাওয়া যায়। এই আইগেন ভেক্টরদ্বয় হলো যথাক্রমে a ও b, যেখানে

$$a = \frac{\sum_{XX}^{-1}\sum_{XY} b}{\sqrt{\lambda}} \qquad \text{(৬.২.৮)}$$

$$b = \frac{\sum_{YY}^{-1}\sum_{YX} a}{\sqrt{\lambda}} \qquad \text{(৬.২.৯)}$$

এখানে a ও b এর মধ্যে যে সম্পর্ক বিদ্যমান তা থেকে বলা যায় যে দুটি নিয়ামক সমীকরণের সমাধান করার প্রয়োজন নেই।

এখন $\sum_{XX}^{-1}\sum_{XY}\sum_{YY}^{-1}\sum_{YX}$ এবং $\sum_{YY}^{-1}\sum_{YX}\sum_{XX}^{-1}\sum_{XY}$

ম্যাট্রিক্সের ক্রমসংখ্যা (rank) হলো q এবং p এর মধ্যে যেটি ন্যূনতম। কিন্তু $q \geqslant p$ হওয়ার কারণে ক্রমসংখ্যা p হবে এবং সর্বোচ্চ p কানুনী চলক (canonical variate) নির্ণয় করা যাবে। ধরা যাক λ এর প্রথম বৃহত্তম মান হলো $\lambda_{(1)}$ এবং এর প্রাসঙ্গিক ভেক্টর হলো $a_{(1)}$ ও $b_{(1)}$। তাহলে প্রথম কানুনী চলক জোড়া হবে $a'_{(1)}X$ এবং $b'_{(1)}Y$। এখানে $a'_{(1)}$ এবং $b'_{(1)}$ হলো কানুনী ভর। যেহেতু $a'\Sigma_{XX}a = b'\Sigma_{YY}b = 1$, সে কারণে $a'_{(1)}$ ও $b'_{(1)}$ হলো আদর্শায়িত চলক X ও Y এর কানুনী ভর।

যেহেতু $\lambda_{(1)}$ হলো প্রথম বৃহত্তম আইগেন মান, সে কারণে কানুনী চলক জোড়া $a'_{(1)}X$ ও $b'_{(1)}Y$ এর কানুনী সংশ্লেষাঙ্ক $\left[\lambda_{(1)}^{1/2}\right]$ বৃহত্তম। এরপর λ এর দ্বিতীয় বৃহত্তম মানের জন্য দ্বিতীয় কানুনী চলক জোড়া নির্ণয় করা যায় যে জোড়ার কানুনী সংশ্লেষাঙ্ক দ্বিতীয় বৃহত্তম হবে এবং ঐ জোড়া প্রথম কানুনী চলক জোড়ার অনপেক্ষ হবে। এভাবে পর্যায়ক্রমে p কানুনী চলক জোড়া নির্ণয় করতে হবে যেন p-তম জোড়া পূর্ববর্তী $(p-1)$ জোড়ার প্রত্যেক জোড়ার অনপেক্ষ হয়। অর্থাৎ Y এর p কানুনী চলকসমূহ পরস্পর অনপেক্ষ হবে। অনুরূপভাবে X এর q কানুনী চলকসমূহ পরস্পর অনপেক্ষ হবে। তাছাড়া X এর j-তম ও Y এর i-তম $(j \neq i)$ কানুনী চলকের মধ্যে কোনো সংশ্লেষণ থাকবে না।

উপরিউক্ত বিশ্লেষণের ক্ষেত্রে ধরা যাক rank $(\Sigma_{XY}) = k$ । সুতরাং Σ_{XY} এর শূন্য নয় এমন আইগেন মান হলো k । মনে করা যাক যে এই k আইগেন মানের সব কয়টি ভিন্ন ভিন্ন এবং $\lambda_1 > \lambda_2 > \cdots > \lambda_k > 0$ । তাহলে, a_i এবং b_i $(i = 1, 2, \cdots, k)$ হলো X এবং Y এর জন্য i-তম কানুনী সংশ্লেষণ ভেক্টর যেখানে $a_i = \sum_{XX}^{-1/2} \alpha_i$, $b_i = \sum_{YY}^{-1/2} \beta_i$ এবং α_i ও β_i হলো যথাক্রমে $N_1 = KK'$ ও $N_2 = K'K$ এর আদর্শায়িত আইগেন ভেক্টর এবং

$$K = \sum_{XX}^{-1/2} \sum_{XY} \sum_{YY}^{-1/2}$$

কাজেই $a_i'X$ এবং $b_i'Y$ হলো i-তম কানুনী সংশ্লেষণ চলক (canonical correlation variable) ।

আলোচিত কানুনী সংশ্লেষণ মাপনী দ্বারা পরিবর্তিত হয় না । ধরা যাক X ও Y-কে আদি বিন্দু ও মাপনী দ্বারা পরিবর্তন করে যথাক্রমে পাওয়া যাচ্ছে

$$X^S = U'X + u \text{ এবং } Y^S = V'Y + v$$

যেখানে U এবং V হলো যথাক্রমে $(q \times q)$ এবং $(p \times p)$ নন্-সিঙ্গুলার ম্যাট্রিক্স এবং u ও v হলো যথাক্রমে $(q \times 1)$ ও $(p \times 1)$ স্থির ভেক্টর । তাহলে

$$V(X^S) = U'\Sigma_{XX}U, \; V(Y^S) = V'\Sigma_{YY}V \text{ এবং}$$

$$Cov(X^S, Y^S) = U'\Sigma_{XY}V$$

সুতরাং X^S ও Y^S এর ক্ষেত্রে

$$\sum_{XX}^{-1} \sum_{XY} \sum_{YY}^{-1} \sum_{YX}$$

এর সমতুল ম্যাট্রিক্স হলো

$$\left(U'\sum_{XX}U\right)^{-1} U'\sum_{XY}V\left(V'\sum_{YX}V\right)^{-1} V'\sum_{YX}U$$

$$= U^{-1}\sum_{XX}^{-1}\sum_{XY}\sum_{YY}^{-1}\sum_{YX}U$$

কিন্তু উক্ত ম্যাট্রিক্সের আইগেন মান হলো

$$\sum_{XX}^{-1}\sum_{XY}\sum_{YY}^{-1}\sum_{XY}$$

এর আইগেন মানের সমান। অনুরূপভাবে

$$V^{-1}\sum_{YY}^{-1}\sum_{YX}\sum_{XX}^{-1}\sum_{XY}V$$

এর আইগেন মান হলো

$$\sum_{YY}^{-1}\sum_{YX}\sum_{XX}^{-1}\sum_{XY}$$

এর আইগেন মানের সমান। কাজেই X ও Y-কে X^S ও Y^S-এ পরিবর্তন করার কারণে কানুনী সংশ্লেষাঙ্কের কোনো পরিবর্তন হয় না। কিন্তু X^S ও Y^S-এ পরিবর্তন করার কারণে আইগেন ভেক্টর দাঁড়ায় $a^S = U^{-1}a$ এবং $b^S = V^{-1}b$। উক্ত পরিবর্তনের কারণে আদর্শায়িতকরণের কোনো পরিবর্তন হয় না। কারণ

$$a^{S'}(U'\Sigma_{XX}U)a^S = a'\Sigma_{XX}a = 1$$

উপরিউক্ত পরিবর্তনের জন্য $U = (\text{diag}\,\Sigma_{XX})^{-1/2}$ এবং

$V = (\text{diag}\,\Sigma_{YY})^{-1/2}$ হয়, তাহলে $\sum_{XX}^{-1}\sum_{XY}\sum_{YY}^{-1}\sum_{YX}$ ম্যাট্রিক্স পরিবর্তিত হয়ে দাঁড়ায় $P_{XX}^{-1}P_{XY}P_{YY}^{-1}P_{YX}$ ম্যাট্রিক্স-এ, এখানে

$$P = \begin{bmatrix} P_{XX} & P_{XY} \\ P_{YX} & P_{YY} \end{bmatrix}$$

হলো X ও Y এর সংশ্লেষাঙ্ক ম্যাট্রিক্স। কাজেই Σ এর পরিবর্তে সংশ্লেষাঙ্ক ম্যাট্রিক্স P ব্যবহার করা হলে কানুনী সংশ্লেষণ বিশ্লেষণের কোনো পরিবর্তন হয় না।

৬.৩ নমুনা কানুনী সংশ্লেষণ বিশ্লেষণ (Sample Canonical Correlation Analysis)

৬.৩.১ বিশ্লেষণ পদ্ধতি (Method of Analysis) : পূর্ববর্তী অনুচ্ছেদে গণসমষ্টি উপাত্ত হতে কানুনী সংশ্লেষণ বিশ্লেষণ পদ্ধতি আলোচনা করা হয়েছে। বাস্তবে

গণসমষ্টি উপাত্ত বিশ্লেষকের গবেষণার সুযোগ-সুবিধার আওতায় নাও থাকতে পারে। সেক্ষেত্রে নমুনা হতে বিশ্লেষণ করাই বেশির ভাগ ক্ষেত্রে প্রচলিত। ধরা যাক X ও Y চলকের নমুনাভিত্তিক Σ_{XX}, Σ_{XY} এবং Σ_{YY} এর নিরূপক হলো যথাক্রমে S_{XX}, S_{XY} এবং S_{YY}। তাহলে গণসমষ্টি উপাত্ত হতে বিশ্লেষণের জন্য যে দুটি ম্যাট্রিক্সের আইগেন মান কানুনী সংশ্লেষাঙ্ক সরবরাহ করে নমুনাভিত্তিক ঐ দুটি ম্যাট্রিক্স হলো

$$S_{XX}^{-1} S_{XY} S_{YY}^{-1} S_{YX} \quad \text{এবং} \quad S_{YY}^{-1} S_{YX} S_{XX}^{-1} S_{XY}$$

কাজেই বিশ্লেষণের বাকি সব ধাপ পূর্ববর্তী অনুচ্ছেদের ন্যায় হবে। আবার X ও Y-কে আদর্শায়িত করা হলে S_{XX}, S_{XY} এবং S_{YY} ম্যাট্রিক্সত্রয় যথাক্রমে R_{XX}, R_{XY} এবং R_{YY} ম্যাট্রিক্স-এ পরিণত হয়। এখানে R দ্বারা সংশ্লেষাঙ্ক ম্যাট্রিক্স বুঝানো হয়েছে। সুতরাং কানুনী সংশ্লেষাঙ্ক এবং কানুনী ভরের নিরূপক পাওয়ার জন্য

$$R_{XX}^{-1} R_{XY} R_{YY}^{-1} R_{YX} \quad \text{এবং} \quad R_{YY}^{-1} R_{YX} R_{XX}^{-1} R_{XY}$$

ম্যাট্রিক্সের আইগেন মান ও আইগেন ভেক্টর নির্ণয় করতে হবে।

ধরা যাক নমুনা হতে প্রাপ্ত a ও b এর মান হলো যথাক্রমে $\hat{a}$ এবং $\hat{b}$। মনে করা যাক X ও Y বহুচলক পরিমিত বিন্যাস থেকে চয়ন করা নমুনা। তাহলে S_{XX}, S_{YY} এবং S_{XY} যথাক্রমে Σ_{XX}, Σ_{YY} এবং Σ_{XY} এর সর্বোত্তম সম্ভাব্য নিরূপক। সুতরাং নমুনা কানুনী সংশ্লেষণের বিভিন্ন মান গণসমষ্টি কানুনী সংশ্লেষণের মানসমূহের সর্বোত্তম সম্ভাব্য নিরূপকসমূহ। অবশ্য নিরূপকসমূহের এই ধর্ম বহাল থাকার জন্য আইগেন মানসমূহ ভিন্ন ভিন্ন হতে হবে।

নমুনা ভেক্টর $\hat{a}$ ও $\hat{b}$ ভেদাঙ্ক ম্যাট্রিক্স হতে নির্ণয় করা হলে এগুলো চলকসমূহের এককের সমানুপাতিক হবে এবং কানুনী চলকসমূহের প্রসারতা (Dimensionality) অর্থবহ হবে। কিন্তু সংশ্লেষাঙ্ক ম্যাট্রিক্স হতে বিশ্লেষণ করা হলে কানুনী চলকসমূহের কোনো প্রসারতা থাকে না। কানুনী সাফল্যাঙ্ক (canonical scores) নির্ণয় করার সময় বিষয়টি স্মরণ রাখা প্রয়োজন।

৬.৩.২ যাচাই পদ্ধতি (Method of Test) : যাচাই পদ্ধতি দুটি বিষয় চিন্তা করতে হয়। প্রথমত, যে চলকসমূহ নিয়ে কানুনী সংশ্লেষণ করা হবে ঐগুলোর মধ্যে কোনো সংশ্লেষণ আছে কিনা ? অর্থাৎ যাচাই করে দেখতে হবে

$$H_0 : \Sigma_{XY} = 0 \qquad (৬.৩.১)$$

সত্য কিনা। কারণ X ও Y অসংশ্লেষিত হলে Σ_{XY} শূন্য হবে এবং সেক্ষেত্রে কানুনী সংশ্লেষণের প্রয়োজন নেই। দ্বিতীয়ত, কানুনী সংশ্লেষণ করা যুক্তিসঙ্গত হলেও কয়টি কানুনী চলক তাৎপর্যপূর্ণ তা যাচাই করে দেখতে হয়।

নাস্তিকল্পনা ৬.৩.১ যাচাই করার জন্য Bartlett (1951) একটি যাচাই পদ্ধতি আলোচনা করেছেন। তাঁর প্রস্তাব অনুযায়ী যাচাই নির্দেশক হলো Wilk's $\Lambda(p, n-1-q, q)$, এখানে

$$\Lambda = \prod_{i=1}^{k} (1-\hat{\lambda}_{(i)}) = \frac{\left| S_{YY}^{-1} S_{YX} S_{XX}^{-1} S_{XY} \right|}{\left| S_{YY} \right|} \qquad (৬.৩.২)$$

$k = \min(p, q)$, $\hat{\lambda}_{(i)}$ হলো $S_{YY}^{-1} S_{YX} S_{XX}^{-1} S_{XY}$-এর i-তম আইগেন মান। Bartlett দেখিয়েছেন যে n বড় হলে Λ pq স্বাধীনতার মাত্রাসহ কাইবর্গ বিন্যাস অনুসরণ করে। এখানে

$$\chi^2 = -\left[(n-1) - \tfrac{1}{2}(q+p+1)\right] \ln \Lambda \qquad (৬.৩.৩)$$

এই $\chi^2 > \chi^2_\alpha$ হলে নাস্তিকল্পনা বাতিল বলে পরিগণিত হবে এবং কানুনী সংশ্লেষণ বিশ্লেষণ যুক্তিসঙ্গত বলে বিবেচিত হবে।

নাস্তিকল্পনা ৬.৩.১ বাতিল হওয়ার পর প্রথম কানুনী চলক জোড়ার প্রভাব Λ হতে বাদ দিয়ে বাকি চলক জোড়াসমূহের তাৎপর্য যাচাই করে দেখা প্রয়োজন। সেক্ষেত্রে

$$\Lambda_1 = \prod_{i=2}^{k} (1-\hat{\lambda}_{(i)})$$

এবং $$\chi_1^2 = -\left[(n-1) - \tfrac{1}{2}(q+p+1)\right] \ln \Lambda_1 \qquad (৬.৩.৪)$$

এই χ_1^2 এর বিন্যাস হলো $(q-1)(p-1)$ স্বাধীনতার মাত্রাবিশিষ্ট কাইবর্গ বিন্যাস। এ পর্যায়ে নাস্তিকল্পনার বিপক্ষে কোনো যুক্তি না থাকলে এক জোড়া কানুনী চলকই নমুনা উপাত্তের জন্য পর্যাপ্ত বলে বিবেচিত হবে। অন্যথায় Λ থেকে প্রথম দুটি কানুনী চলক জোড়ার প্রভাব বাদ দিয়ে আবার নাস্তিকল্পনা যাচাই করতে হবে। সাধারণত প্রথম $k' < k = \min(q, p)$ কানুনী চলকের প্রভাব বাদ

দিয়ে বাকি চলক জোড়াসমূহের তাৎপর্য যাচাই করতে হলে পরিবর্তিত Λ-এর মান হবে

$$\Lambda^* = \prod_{i=k'+1}^{k} (1-\hat{\lambda}_{(i)}) \qquad (৬.৩.৫)$$

এবং যাচাই তথ্যজমান হবে

$$\chi^{*2} = -\left[(n-1)-\tfrac{1}{2}(q+p+1)\right]\ln\Lambda^* \qquad (৬.৩.৬)$$

এই χ^{*2} এর বিন্যাস হলো $(q-k')(p-k')$ স্বাধীনতার মাত্রাবিশিষ্ট কাইবর্গ বিন্যাস।

উপরিউক্ত যাচাই পদ্ধতি প্রয়োগ করার ক্ষেত্রে সতর্কতার প্রয়োজন। প্রথম কানুনী চলক জোড়ার তাৎপর্য যাচাই-এ কোনো সন্দেহের সুযোগ নেই। কিন্তু প্রথমটির পরবর্তী যে কোনো চলক জোড়ার সংশ্লেষাঙ্ক 1 এর কাছাকাছি না হলে বুঝতে হবে যে আরো তাৎপর্যপূর্ণ কানুনী চলক জোড়াসমূহ থাকতে পারে।

আগেই উল্লেখ করা হয়েছে যে কানুনী সংশ্লেষণ বিশ্লেষণের দ্বারা দুই গুচ্ছ চলকের রৈখিক সমাবেশের সংশ্লেষণ পর্যালোচনা করা হয়। আদি চলক গুচ্ছের কোনো সংশ্লেষণ পর্যালোচনা করতে এটি সাহায্য করে না। ফলে $k' < k = \min(q, p)$ সংখ্যক তাৎপর্য চলক জোড়া পেলেও ঐগুলো আদি চলক গুচ্ছসমূহের জন্য তেমন গুরুত্বপূর্ণ কোনো তথ্য সরবরাহ নাও করতে পারে। সে কারণে শুধু যাচাই নির্দেশকের উপর নির্ভর না করে বিশ্লেষণের পর্যাপ্ততার জন্য অন্যান্য নির্দেশকও পর্যালোচনা করা প্রয়োজন। এ সম্পর্কে পরবর্তী অনুচ্ছেদে আলোচনা করা হবে।

৬.৪ কানুনী সংশ্লেষণ বিশ্লেষণ হতে তাৎপর্য নির্ণয় (Interpretation from Canonical Correlation Analysis)

আগেই আলোচনা করা হয়েছে যে কানুনী সংশ্লেষণ বিশ্লেষণের ক্ষেত্রে একগুচ্ছ নির্ভরশীল চলকের একটি রৈখিক সমাবেশ এবং একগুচ্ছ অনপেক্ষ চলকের একটি রৈখিক সমাবেশ এমনভাবে নির্ণয় করতে হয় যেন দুটি রৈখিক সমাবেশের সংশ্লেষণ বৃহত্তম হয়। আবার বহুচলক নির্ভরণের ক্ষেত্রে একগুচ্ছ অনপেক্ষ চলকের সাথে একটি নির্ভরশীল চলকের সংশ্লেষণ পর্যালোচনা করা হয়। ফলে কানুনী সংশ্লেষণের ক্ষেত্রে নির্ভরশীল চলকের গুচ্ছে একটি চলক হলে, ঐ বিশ্লেষণ বহুচলক নির্ভরণের ন্যায় হয়। কাজেই বহুচলক নির্ভরণের ন্যায় কানুনী সংশ্লেষণ বিশ্লেষণ হতে প্রাপ্ত ফলাফলের তাৎপর্যও নির্ণয় করার প্রশ্ন দাঁড়ায়। বর্তমান অনুচ্ছেদে এই তাৎপর্য ব্যাখ্যা করা হবে। এখানে একটি কথা মনে রাখতে হবে যে নির্ভরশীল চলক

গুচ্ছের একটি করে নিয়ে অনপেক্ষ চলক গুচ্ছের সাথে নির্ভরণ বিশ্লেষণ করা হলে কানুনী সংশ্লেষণ বিশ্লেষণের ন্যায় ফলাফল পাওয়া যাবে না। কারণ ভিন্ন ভিন্ন বহুচলক নির্ভরণের ক্ষেত্রে নির্ভরশীল চলক গুচ্ছের মধ্যে যে আন্তঃসম্পর্ক বিদ্যমান তা বিবেচনা করা হয় না। তাহলে প্রশ্ন হলো কানুনী সংশ্লেষণ বিশ্লেষণ ও বহু-চলক নির্ভরণ বিশ্লেষণের মধ্যে সঠিক সম্পর্ক কি? নিচে এ সম্পর্ক ব্যাখ্যা করা যাক।

৬.৪.১ **বহু-নির্ভরণ বিশ্লেষণের সাথে সম্পর্ক** (Relation to Multiple Regression Analysis) : ধরা যাক $q=p=1$, তাহলে সংশ্লেষণ ম্যাট্রিক্স হতে কানুনী সংশ্লেষণ বিশ্লেষণ করার ক্ষেত্রে পাওয়া যায়

$$R_{YY}^{-1} R_{YX} R_{XX}^{-1} R_{XY} = R_{XX}^{-1} R_{XY} R_{YY}^{-1} R_{YX} = R_{YX} R_{XY} = r^2$$

এখানে r^2 হলো X ও Y এর সংশ্লেষাঙ্কের বর্গ।

বহু-নির্ভরণের ক্ষেত্রে q অনপেক্ষ চলক ও $p=1$ নির্ভরশীল চলক হতে পাওয়া যায়

$$\hat{b} = R_{XX}^{-1} R_{XY}$$

আবার $$R_{YY}^{-1} R_{YX} R_{XX}^{-1} R_{XY} = R_{YX} R_{XX}^{-1} R_{XY} = R_{YX} \hat{b}$$

এখানে $\hat{b}$ হলো আদর্শায়িত আংশিক নির্ভরাঙ্কসমূহের $(q \times 1)$ স্তম্ভ ভেক্টর। এখানে $R_{XY} \hat{b}$ হলো স্কেলার এবং বহু-নির্ভরণের ক্ষেত্রে এটি হলো বহু-সংশ্লেষাঙ্কের বর্গ। কাজেই একটি নির্ভরশীল চলকের ক্ষেত্রে কানুনী সংশ্লেষণ বিশ্লেষণ বহু-নির্ভরণ বিশ্লেষণের সমতুল।

৬.৪.২ **কানুনী ভার** (Canonical Loading) : আগেই উল্লেখ করা হয়েছে যে, কানুনী ভর (canonical weight) হলো কানুনী চলক পাওয়ার জন্য আদি চলকের গুরুত্ব নির্দেশক একটি পরিমাপ। সুতরাং ভরগুলো বহু-নির্ভরাঙ্ক বিশ্লেষণের ক্ষেত্রে আংশিক নির্ভরাঙ্কসমূহের সাথে তুলনীয়। আসলে, এগুলো বহু-নির্ভরাঙ্ক বিশ্লেষণের ক্ষেত্রে β-সহগের সমতুল এবং এগুলো কানুনী চলকের ভেদে আদি চলকসমূহের অবদান নির্দেশ করে। কিন্তু আদি চলকসমূহের মধ্যে মাল্টিকোলিনিয়ারিটি বিদ্যমান থাকলে কানুনী ভর দ্বারা আদি চলকসমূহের সাথে কানুনী চলকের সঠিক সম্পর্ক নির্দেশিত হয় না। তাছাড়া মাল্টিকোলিনিয়ারিটির

দ্বিতীয় ম্যাট্রিক্সের আইগেন ভেক্টর হলো

$$\hat{b} = \begin{matrix} \\ \\ \\ \end{matrix} \begin{matrix} b_1 \\ \\ \\ \end{matrix} \begin{matrix} b_2 \\ \\ \\ \end{matrix} \begin{matrix} b_3 \\ \\ \\ \end{matrix} \begin{bmatrix} 1.11866 & -0.29079 & 0.58747 \\ -0.20158 & -0.06718 & -1.20479 \\ -0.07064 & 1.04421 & -0.30222 \end{bmatrix}$$

উক্ত উপাত্তের জন্য $\Lambda = 0.52231$, $q = p = 3$, $n = 57$ । কাজেই ৬.৩.৩ সমীকরণ অনুসারে $\chi^2 = 34.10$ । এই χ^2 এর স্বাধীনতার মাত্রা হলো q । এখানে $p(\chi^2 \geq 34.10) = 0.0001$ হওয়াতে নাস্তিকল্পনা $H_0 : P_{XY} = 0$ বাতিল বলে পরিগণিত হলো এবং আলোচিত চলকসমূহের ভিত্তিতে কানুনী সংশ্লেষণ বিশ্লেষণ করা যুক্তিসঙ্গত মনে করা যায় । আবার, যাচাই তথ্যজমান ৬.৩.৪ অনুসারে পাওয়া যায় $p(\chi^2 \geq 8.97) = 0.0618$ । এই χ^2 এর স্বাধীনতার মাত্রা হলো 4 । কাজেই প্রথম জোড়া কানুনী চলকই আলোচিত উপাত্তের জন্য পর্যাপ্ত । এখানে প্রথম কানুনী চলক জোড়া হলো

$$a'_{(1)} X = 0.90225\,A - 0.26307\,B + 0.11961\,E$$

এবং $$b'_{(1)} Y = 1.11866\,G - 0.20158\,H - 0.07064\,I$$

এই চলক জোড়ার জন্য কানুনী ভর হলো

$$a'_{(1)} = [\,0.90225 \quad -0.26307 \quad 0.11961\,] \text{ এবং}$$

$$b'_{(1)} = [\,1.11866 \quad -0.20158 \quad -0.07064\,]$$

আবার কানুনী সংশ্লেষাঙ্ক হলো $\lambda_{(1)}^{1/2} = 0.6167$ । এই বিশ্লেষণের বিস্তারিত ফলাফল সারণি ৬.১-এ দেয়া হলো ।

সারণি ৬.১ : কানুনী সংশ্লেষণ বিশ্লেষণের ফলাফল ।

চলক সংখ্যা	আইগেন মান, λ	কানুনী সংশ্লেষাঙ্ক, $\sqrt{\lambda}$	Λ	χ^2	স্বাধীনতার মাত্রা	p-value
1	0.3803	0.6167	0.5223	34.10	9	0.0001
2	0.1483	0.3851	0.8429	8.97	4	0.0618
3	0.0104	0.1018	0.9896	0.55	1	0.4598

	Y-গুচ্ছের জন্য কানুনী ভর কানুনী চলক			Y-গুচ্ছের জন্য কানুনী ভার কানুনী চলক		
	1	2	3	1	2	3
G	1.1187	−0.2908	0.5875	0.9847	0.0358	−0.1713
H	−0.2016	−0.0672	−1.2048	0.4063	−0.2384	−0.8821
I	−0.0706	1.0442	−0.3022	0.3209	0.9440	−0.1135
			ভেদাঙ্কের %	41.2	31.6	27.3

	X-গুচ্ছের জন্য কানুনী ভর কানুনী চলক			X-গুচ্ছের জন্য কানুনী ভার কানুনী চলক		
	1	2	3	1	2	3
A	0.9023	0.4595	−0.4098	0.9795	0.0945	−0.1777
B	−0.2631	0.5015	−1.1840	−0.5556	0.7197	−0.4164
E	0.1196	0.6845	1.0237	−0.2506	0.8702	0.4243
			ভেদাঙ্কের %	44.4	42.8	12.8

উপরিউক্ত উপাত্তের ক্ষেত্রে লক্ষ্য করা যাচ্ছে যে মোট জীবিত জন্মগ্রহণ করা সন্তানের সংখ্যা (G) এবং মোট মৃত সন্তানের সংখ্যা (H) ধনাত্মকভাবে এবং তাৎপর্যপূর্ণভাবে সংশ্লেষিত। অর্থনৈতিক অবস্থার (I) উন্নতির সাথে সাথে সন্তানের সংখ্যা তাৎপর্যপূর্ণভাবে বৃদ্ধি পেয়েছে। কিন্তু আর্থিক অবস্থা শিশু মৃত্যুর উপর কোনো প্রভাব ফেলতে পারেনি [R_{YY} দ্রষ্টব্য]। আবার, R_{XX} হতে লক্ষ্য করা যাচ্ছে যে অল্প বয়সের মায়েদের শিক্ষা বেশি এবং অধিকাংশ শিক্ষিত স্বামীর স্ত্রীরাও শিক্ষিত। শিক্ষিত মায়েদের সন্তান সংখ্যাও কম [R_{YX} ম্যাট্রিক্স]। কিন্তু মায়ের শিক্ষা বা বয়স কোনোটিই শিশু মৃত্যুর উপর তাৎপর্যপূর্ণ প্রভাব ফেলতে পারেনি।

গুচ্ছ দ্বারা Y-চলক গুচ্ছের ভেদের ব্যাখ্যা করা সমানুপাত বা X এর ভিত্তিতে Y এর Redundancy co-efficiency $\left(R^2_{(i)Y/X}\right)$ হলো

$$R^2_{(i)Y(X)} = \lambda_i \sum_{j=1}^{p} \frac{\left(r_{Y^*Y_j(i)}\right)^2}{p} = \lambda_i R^2_{(i)Y}$$

Stewart and Love দেখিয়েছেন যে Y চলক গুচ্ছের ভিত্তিতে ভিন্ন ভিন্ন নির্ভরণ বিশ্লেষণ করা হলে প্রতি নির্ভরণ হতে প্রাপ্ত বহু-সংশ্লেষাঙ্ক-এর বর্গসমূহের গড় নেয়া হলে তা Redundancy co-efficient এর সমতুল হয়। এ সম্পর্কে আরো বিস্তারিত জানার জন্য Miller (1975), Gleason (1976), Wollenberg (1977) পর্যালোচনা করা যেতে পারে।

৬.৫ সাফল্যাঙ্ক এবং পূর্বাভাস (Score and Prediction)

ইতোমধ্যেই আলোচনা করা হয়েছে যে অনপেক্ষ চলক গুচ্ছ X ও নির্ভরশীল চলক গুচ্ছ Y এর কানুনী সংশ্লেষণ বিশ্লেষণ এমনভাবে করা হয় যেন $X^* = a'X$ এবং $Y^* = b'Y$ এর সংশ্লেষাঙ্ক বৃহত্তম হয়। অর্থবহ বিশ্লেষণের ক্ষেত্রে যে কোনো কানুনী চলক X-গুচ্ছ ও Y-গুচ্ছ চলকের বৃহত্তর অংশ ব্যাখ্যা করার কথা। বিশ্লেষণের সফলতা নির্ভর করে চলক গুচ্ছের ব্যাখ্যা করা ভেদের উপর। এখানে বিশ্লেষণের সফলতা নিয়ে আলোচনা করা হবে।

ধরা যাক a_i এবং b_i $(i = 1, 2, \ldots, k)$ হলো i-তম কানুনী সংশ্লেষণ ভেক্টর। তাহলে, i-তম কানুনী সংশ্লেষণ ভেক্টর হতে X ও Y-চলক গুচ্ছের সাফল্যাঙ্ক দাঁড়ায় যথাক্রমে Xa_i এবং Yb_i। এই সাফল্যাঙ্কের ভিত্তিতে X এর এবং Y এর n নমুনা বিন্দুর মান হলো যথাক্রমে $X_i^* = a_i'X$ এবং $Y_i^* = b_i'Y$

$$[i = 1, 2, \ldots, k = \min(p, q)]$$

এখন X ও Y চলক যথাক্রমে অনপেক্ষ ও নির্ভরশীল চলক হিসেবে বিবেচিত হলে সাফলাঙ্ক X_i^* ব্যবহার করে সাফলাঙ্ক Y_i^* এর পূর্বাভাস করা যায়। ন্যূনতম বর্গ পদ্ধতির নির্ভরণ বিশ্লেষণের মাধ্যমে পূর্বাভাস করা হয় এবং Y_i^* এর পূর্বাভাসকৃত মান হলো

$$\hat{Y}_i^* = \sqrt{\lambda_i}\,(X_i^* - a_i'\bar{X}_i) + b_i'\bar{Y}$$

এখানে $\sqrt{\lambda_i}$ হলো i-তম কানুনী সংশ্লেষাঙ্ক। উল্লেখযোগ্য যে, X এর উপর নির্ভরণ করার কারণে Y_i^* এর ভেদের ব্যাখ্যা করা সমানুপাত হলো λ_i।

উদাহরণ ৬.১ ঃ উদাহরণ ৫.১ থেকে চলক {A, B, E} গুচ্ছকে অনপেক্ষ চলক বিবেচনা করে এবং চলক {G, H, I} গুচ্ছকে নির্ভরশীল চলক বিবেচনা করে কানুনী সংশ্লেষণ বিশ্লেষণ করা যাক।

উপরিউক্ত উপাত্তের জন্য নমুনা সংশ্লেষাঙ্ক ম্যাট্রিক্সগুলো হলো

$$R_{XX} = \begin{array}{c} \\ A \\ B \\ E \end{array} \begin{array}{c} \begin{array}{ccc} A & B & E \end{array} \\ \begin{bmatrix} 1.0000 & & \\ -0.4023^{**} & 1.0000 & \\ -0.2386 & 0.5888^{**} & 1.0000 \end{bmatrix} \end{array},$$

$$R_{YY} = \begin{array}{c} \\ G \\ H \\ I \end{array} \begin{array}{c} \begin{array}{ccc} G & H & I \end{array} \\ \begin{bmatrix} 1.0000 & & \\ 0.5426^{**} & 1.0000 & \\ 0.3477^{**} & -0.0129 & 1.0000 \end{bmatrix} \end{array}$$

এবং

$$R_{XY} = \begin{array}{c} \\ A \\ B \\ E \end{array} \begin{array}{c} \begin{array}{ccc} G & H & I \end{array} \\ \begin{bmatrix} 0.5993^{**} & 0.2527 & 0.2105 \\ -0.3257^{**} & -0.1649 & 0.0749 \\ -0.1533 & -0.1905 & 0.2113 \end{bmatrix} \end{array} = R'_{YX}$$

$^{*}\Rightarrow p-\text{value} < 0.01$, $^{**}\Rightarrow p-\text{value} < 0.001$

এখানে $R_{XX}^{-1} R_{XY} R_{YY}^{-1} R_{YX}$ বা $R_{YY}^{-1} R_{YX} R_{XX}^{-1} R_{XY}$ এর

আইগেন মানসমূহ হলো $\lambda_{(1)} = 0.3803 > \lambda_{(2)} = 0.1483 > \lambda_{(3)} = 0.0104$ ।

প্রথমোক্ত ম্যাট্রিক্সের আইগেন ভেক্টর হলো

$$a = \begin{array}{c} \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} a_1 & a_2 & a_3 \end{array} \\ \begin{bmatrix} 0.90225 & 0.45954 & -0.40977 \\ -0.26307 & 0.50149 & -1.18396 \\ 0.11961 & 0.68453 & 1.02366 \end{bmatrix} \end{array}$$

এর কারণে ভরগুলোর চিহ্ন পরিবর্তিত হয়ে যেতে পারে। কারণ কোনো কোনো চলকের ভেদাঙ্ক অন্য চলক দ্বারা প্রকাশিত হয়ে থাকতে পারে। এই একই কারণে কানুনী ভর ছোটও হয়ে থাকতে পারে, আবার চলক পরিবর্তনের সাথে সাথে ভরও পরিবর্তিত হয়ে থাকতে পারে। এ সব কারণে কানুনী ভর (canonial weights) এর পরিবর্তে কানুনী ভার (canonical loading) নির্ণয় করা হয় এবং একই ভার কানুনী সংশ্লেষণ পর্যালোচনা করার জন্য ব্যবহার করা হয়।

কানুনী ভার হলো আদি চলকের সাথে তার প্রাসঙ্গিক কানুনী চলকের সরল সংশ্লেষণ পরিমাপ করার একটি নির্দেশক। কানুনী চলক দ্বারা কোন চলকের কতটুকু ব্যাখ্যা করা যায় তার মাত্রা নির্দেশ করে কানুনী ভার। এই ভার নির্ণয় করার জন্য ধরা যাক X ও Y চলক গুচ্ছের জন্য i-তম কানুনী চলকের ভর হলো যথাক্রমে $\hat{a}_i$ এবং $\hat{b}_i$। আবার X চলক গুচ্ছের জন্য সংশ্লেষণ ম্যাট্রিক্স হলো R_{XX} এবং Y চলক গুচ্ছের জন্য এটি হলো R_{YY}। তাহলে X ও Y চলক গুচ্ছের জন্য কানুনী ভার (canonical loadings) হলো যথাক্রমে

$$R_{X*X(i)} = R_{YY}\,\hat{a}_i$$

এবং

$$R_{Y*Y(i)} = R_{YY}\,\hat{b}_i$$

অবশ্য এই কানুনী ভারও স্থিতিশীল হবে এমন কোনো কথা নেই। তাছাড়া স্থিতিশীল হলেও কানুনী ভার X ও Y চলক গুচ্ছের মধ্যে একটি কাল্পনিক সম্পর্ক স্থাপন করে। কারণ Y গুচ্ছ $\hat{b}'Y$ এর সাথে সম্পর্কিত, X গুচ্ছ $\hat{a}'X$ এর সাথে সম্পর্কিত। আবার $\hat{a}'X$ ও $\hat{b}'Y$ সংশ্লেষিত। কাজেই X ও Y সম্পর্কিত। এখানে X ও Y চলক গুচ্ছ তাদের উচ্চ কানুনী ভার এর কারণে সম্পর্কিত। সুতরাং এ ধরনের সম্পর্ক থেকে কোনো সঠিক সিদ্ধান্ত নেয়া সহজ নয়।

৬.৪.৩ ক্রস ভার (Cross Loading) : কানুনী ভার (canonical loading) হলো আদি চলকের সাথে তার প্রাসঙ্গিক কানুনী চলকের সরল সংশ্লেষণের একটি পরিমাপ। আর ক্রস ভার (cross loading) হলো X চলক গুচ্ছ হতে যে কোনো একটি চলকের সাথে $\hat{b}'Y$ এর বা Y চলক গুচ্ছ হতে যে কোনো একটি চলকের সাথে $\hat{a}'X$ এর সরল সংশ্লেষণ পরিমাপ করার একটি নির্দেশক। এই সংশ্লেষণের পরিমাপ হলো কানুনী ভার এবং কানুনী সংশ্লেষাঙ্ক এর গুণফল। ধরা যাক Y_j এর সাথে X গুচ্ছের i-তম কানুনী চলকের সংশ্লেষাঙ্ক (ক্রস ভার) হলো $r_{X*Y_j(i)}$, এখানে

$$r_{X*Y_{j(i)}} = R_{Y*Y_{j(i)}} X \sqrt{\hat{\lambda}_i}$$

অনুরূপভাবে X_i এর সাথে Y গুচ্ছের i-তম কানুনী চলকের ক্রস ভার হলো

$$r_{Y*X_{j(i)}} = R_{X*X_{j(i)}} X \sqrt{\hat{\lambda}_i}$$

৬.৪.৪ **ব্যাখ্যা করা ভেদের সমানুপাত** (Proportion of Explained Variance) : কানুনী সংশ্লেষণ বিশ্লেষণ আদি নির্ভরশীল চলক গুচ্ছ ও অনপেক্ষ চলক গুচ্ছ-এর এমন দুটি রৈখিক সমাবেশ সরবরাহ করে যে সমাবেশদ্বয়ের সংশ্লেষণ বেশি। প্রথম কানুনী চলক জোড়া সমাবেশদ্বয়ের বৃহত্তম সংশ্লেষাঙ্ক সরবরাহ করে। তাহলে, যে কোনো কানুনী চলক X-গুচ্ছ ও Y-গুচ্ছের ভেদের কত অংশ ব্যাখ্যা করে থাকে তা জানার বিষয়। কারণ আদি চলক গুচ্ছের ভেদের বৃহত্তর অংশ কোনো কানুনী চলক ব্যাখ্যা করতে না পারলে ঐ কানুনী চলক তাৎপর্যহীন। এখানে Y-গুচ্ছ ও X-গুচ্ছ চলকের ভেদের কত অংশ ব্যাখ্যা করে তার পরিমাপ দেয়া হলো।

ধরা যাক i-তম কানুনী চলক দ্বারা Y-গুচ্ছের ভেদের ব্যাখ্যা করা সমানুপাত হলো $R^2_{(i)Y}$, যেখানে

$$R^2_{(i)Y} = \frac{1}{p} R'_{Y*Y(i)} R_{Y*Y(i)} = \frac{1}{p} \sum_{j=1}^{p} \left(r_{Y*Y_{j(i)}} \right)^2$$

অনুরূপভাবে j-তম কানুনী চলক দ্বারা X-গুচ্ছের ভেদের ব্যাখ্যা করা সমানুপাত

$$R^2_{(i)X} = \frac{1}{q} R'_{X*X(i)} R_{X*X(i)} = \frac{1}{q} \sum_{j=1}^{q} \left(r_{X*X_{j(i)}} \right)^2$$

এখানে $r_{X*Y_{j(i)}}$ হলো j-তম নির্ভরশীল চলকের সাথে সম্পর্কিত i-তম কানুনী চলকের জন্য কানুনী ভার (canonical loading)।

উপরে আলোচিত ভেদের সমানুপাত অপেক্ষা অধিক অর্থবহ ভেদের সমানুপাত হলো X চলক দ্বারা ব্যাখ্যা করা Y চলকের ভেদের সমানুপাত বা Y চলক দ্বারা ব্যাখ্যা করা X চলকের সমানুপাত। এ ধরনের ভেদের সমানুপাত নির্ণয় করার একটি পরিমাপ প্রস্তাব করেছেন Stewart and love (1968)। তাঁরা এ পরিমাপের নাম দিয়েছেন Redundancy co-efficient। i-তম কানুনী চলকের ক্ষেত্রে X চলক

সারণি ৬.১ থেকে লক্ষ্য করা যাচ্ছে যে প্রথম কানুনী চলক জোড়া তাৎপর্যপূর্ণভাবে সংশ্লেষিত। দ্বিতীয় কানুনী চলক জোড়া মোটামুটি সংশ্লেষিত হলেও ঐ সংশ্লেষণের কোনো তাৎপর্য নেই। তৃতীয় কানুনী চলক জোড়ার মধ্যে কোনো সংশ্লেষণ নেই। প্রথম চলক জোড়া X-গুচ্ছ ও Y-গুচ্ছ চলকের ভেদের শতকরা প্রায় 38 ভাগ ব্যাখ্যা করেছে।

কানুনী ভর থেকে লক্ষ্য করা যাচ্ছে যে প্রথম চলক জোড়া জীবিত জন্মগ্রহণ করা সন্তান সংখ্যা (G) এবং মায়ের বয়স (A) দ্বারা বেশি প্রভাবান্বিত। লক্ষ্য করলে দেখা যাবে যে A এবং G এর সরল সংশ্লেষাঙ্কও বড় এবং তাৎপর্যপূর্ণ। দ্বিতীয় কানুনী চলক জোড়া অর্থনৈতিক অবস্থা (I) এবং পিতার শিক্ষা (E) দ্বারা বেশি প্রভাবান্বিত। তৃতীয় কানুনী চলক জোড়া Y-গুচ্ছের G ও H দ্বারা এবং X-গুচ্ছের B এবং E দ্বারা বেশি প্রভাবান্বিত। উভয় ক্ষেত্রেই চলকগুলোর প্রভাব বিপরীতমুখী।

কানুনী ভার থেকে লক্ষ্য করা যাচ্ছে যে জীবিত জন্মগ্রহণ করা সন্তান সংখ্যা (G) Y-গুচ্ছের (G, H, I) কানুনী চলকের সাথে বেশি সংশ্লেষিত এবং মায়ের বয়স (A) X-গুচ্ছের প্রথম কানুনী চলকের সাথে বেশি সংশ্লেষিত। অবশ্য মায়ের শিক্ষাও (B) প্রথম কানুনী চলকের সাথে বেশি সংশ্লেষিত। এই বিশ্লেষণ হতে বলা যায় যে, জীবিত জন্মগ্রহণ করা মোট সন্তান সংখ্যা মায়ের বয়স বাড়ার সাথে বাড়ে এবং মায়ের শিক্ষার স্তর বৃদ্ধির সাথে কমে। দ্বিতীয় কানুনী চলক জোড়া থেকে লক্ষ্য করার বিষয় হলো মা-বাবার শিক্ষার স্তর বৃদ্ধির সাথে সাথে আর্থিক অবস্থার উন্নতি হয়ে থাকে।

এই বিশ্লেষণ হতে আরো লক্ষণীয় বিষয় হলো যে প্রথম কানুনী চলক X-গুচ্ছের চলকসমূহের ভেদের প্রায় 44 শতাংশ এবং Y-গুচ্ছের চলকসমূহের ভেদের প্রায় 41 শতাংশ ব্যাখ্যা করেছে। এখানে

$$R^2_{(1)Y} = \frac{1}{3}\left[(0.9847)^2 + (0.4063)^2 + (0.3209)^2\right] = 0.412$$

$$R^2_{(1)X} = \frac{1}{3}\left[(0.9795)^2 + (-0.5556)^2 + (-0.2506)^2\right] = 0.444$$

উপরিউক্ত বিশ্লেষণ থেকে G এর সাথে X-গুচ্ছের প্রথম কানুনী চলকের ($a_1'X$) সম্পর্ক নির্ণয় করার জন্য ক্রস ভার নিম্নরূপভাবে নির্ণয় করা যায়। এখানে

G এর জন্য ক্রস ভার = G এর কানুনী ভার X কানুনী সংশ্লেষাঙ্ক

$$= 0.9847 \times 0.6167$$

$$= 0.6073$$

অন্যান্য চলকের ক্ষেত্রে ক্রস ভার সারণি ৬.২-এ উপস্থাপন করা হলো ।

সারণি ৬.২ ঃ প্রথম কানুনী চলকের ক্ষেত্রে চলকসমূহের ক্রস ভার ।

কানুনী সংশ্লেষাঙ্ক = 0.6167

চলক	কানুনী ভার	ক্রস ভার
	কানুনী চলক-১	কানুনী চলক-১
G	0.9847	0.6073
H	0.4063	0.2506
I	0.3209	0.1979
A	0.9795	0.6041
B	−0.5556	−0.3426
E	−0.2506	−0.1545

এখানে লক্ষ্য করার বিষয় হলো যে জীবিত জন্মগ্রহণ করা সন্তানের মোট সংখ্যা (G) মায়ের বয়স (A), মায়ের শিক্ষা (B) এবং পিতার শিক্ষা (E) এর রৈখিক সমাবেশের সাথে তাৎপর্যপূর্ণভাবে সংশ্লেষিত । মৃত সন্তানের সংখ্যা (H) এবং অর্থনৈতিক অবস্থা (I) {A, B, E} চলক গুচ্ছের রৈখিক সমাবেশের সাথে খুব বেশি সংশ্লেষিত নয় ।

আলোচিত উদাহরণের ক্ষেত্রে লক্ষণীয় বিষয় হলো যে, জীবিত জন্মগ্রহণ করা সন্তানের সংখ্যা (G), মৃত সন্তানের সংখ্যা (H) এবং অর্থনৈতিক অবস্থা (I)-এর ভেদের শতকরা প্রায় 41 ভাগ [$R^2_{(1)Y} = 0.412$] প্রথম কানুনী চলক [$a'Y$] দ্বারা প্রকাশিত হয়েছে । বাস্তবে মা-বাবার শিক্ষা [B, E] এবং মায়ের বয়স [A]

{G, H, I}-এর ভেদের কত অংশ ব্যাখ্যা করতে পারে তা জানাও আবশ্যক। এটি জানার জন্য Redundancy সহগ নির্ণয় করতে হয়। সারণি ৬.৩-এ Redundancy সহগ দেয়া হলো। লক্ষ্য করা যাচ্ছে যে, {A, B, E} চলক গুচ্ছ দ্বারা

সারণি ৬.৩ ঃ চলকসমূহের Redundancy সহগ।

চলকসমূহ এর গুচ্ছ (1)	কানুনী চলক i (2)	কানুনী সংশ্লেষাঙ্ক এর বর্গ, λ_i (3)	$R^2_{(i)Y}$ (4)	$R^2_{(i)X}$	Redundancy সহগ, $\hat{\lambda}_i(4)$
G, H, I	1	0.3803	0.412		0.157
	2	0.1483	0.316		0.047
	3	0.0104	0.273		0.003
A, B, E	1	0.3803		0.444	0.169
	2	0.1483		0.428	0.063
	3	0.0104		0.128	0.001

{G, H, I} চলক গুচ্ছের প্রায় শতকরা 21 ভাগ ভেদ ব্যাখ্যা করেছে [0.157 + 0.047 + 0.003 = 0.207]। {A, B, E} গুচ্ছের প্রথম কানুনী চলক দ্বারা {G, H, I} গুচ্ছের ভেদের ব্যাখ্যা করা সমানুপাত হলো 0.157। {A, B, E} গুচ্ছের ভেদের প্রায় 17 ভাগ {G, H, I} চলক গুচ্ছ দ্বারা প্রকাশিত হয়েছে।

উপরিউক্ত বিশ্লেষণের ক্ষেত্রে প্রথম কানুনী চলক

$$X_1^* = a'_{(1)}X = 0.90225\,A - 0.26307\,B + 0.11961\,E$$

{A, B, E} চলক গুচ্ছের ভেদের শতকরা প্রায় 44 ভাগ ব্যাখ্যা করেছে। এখানে প্রথম কানুনী চলক জোড়ার ভিত্তিতে প্রতিটি দম্পতির জন্য সাফল্যাঙ্ক [scores, Xa_1 এবং Yb_1] নির্ণয় করা যায়। আবার X_1^*-এর ভিত্তিতে Y_1^* এর মান পূর্বাভাস করা যায়। Y_1^* এর পূর্বাভাসকৃত মান পাওয়ার সূত্র হলো

$$Y_1^* = \sqrt{\hat{\lambda}_1}\,(X_1^* - a_1'\bar{X}) + b_1'\bar{Y}$$

এখানে

$$\bar{X} = \begin{bmatrix} \bar{A} \\ \bar{B} \\ \bar{E} \end{bmatrix} = \begin{bmatrix} 40.84 \\ 1.44 \\ 2.26 \end{bmatrix}, \quad \bar{Y} = \begin{bmatrix} \bar{G} \\ \bar{H} \\ \bar{I} \end{bmatrix} = \begin{bmatrix} 8.46 \\ 0.70 \\ 2.21 \end{bmatrix}$$

তাহলে $a_1' \bar{X} = 36.74$, $b_1' \bar{Y} = 9.17$

$$\therefore \quad Y_1^* = 0.6167\,[0.90225A - 0.26307B + 0.11961E - 36.74] + 9.17$$

$$= 0.56A - 0.16B + 0.07E - 13.49$$

সপ্তম অধ্যায়

গুচ্ছ বিশ্লেষণ
(Cluster Analysis)

৭.১ সূচনা

একই নমুনা বিন্দু হতে অনেকগুলো চলকের পরিমাপ এবং এই পরিমাপ অনেকগুলো নমুনা বিন্দু হতে করা হলে উপাত্তের বিশ্লেষণ বহুচলক বিশ্লেষণের অন্তর্ভুক্ত হয়। এ জাতীয় বিশ্লেষণের ক্ষেত্রে নমুনা বিন্দুসমূহ কোনো একটি বৈশিষ্ট্যের ভিত্তিতে একই জাতীয় (Homogeneous) বলে অনুমান করা হয়। একই গণসমষ্টি হতে নমুনা চয়ন করা হলে নমুনা বিন্দুসমূহের মধ্যে কোনো ধারাবাহিক পার্থক্য (systematic difference) থাকারও কথা নয়। অবশ্য গণসমষ্টির এককসমূহ (units) বিশেষ বৈশিষ্ট্যের ভিত্তিতে একই জাতীয় হবেই এমন কোনো কথা নয়। যেমন, আর্থিক সচ্ছলতার বিবেচনায় কোন দেশের সব লোক একই শ্রেণিভুক্ত, ধনী, গরীব, মধ্যবিত্ত, নিম্নবিত্ত ইত্যাদি শ্রেণিতে জনগণকে বিভক্ত করা যায়। আবার সব শ্রেণির আর্থিক সচ্ছলতার ক্ষেত্রে একই চলক গুচ্ছের প্রভাব থাকে। তবে চলক গুচ্ছের মানের তারতম্য অবশ্যই থাকে। যেমন, উচ্চ শিক্ষিত লোক সাধারণত উচ্চ পদে নিয়োজিত থাকে এবং বেশি বেতন পায়। এক্ষেত্রে শিক্ষার সাথে পেশার এবং আয়ের একটি সম্পর্ক আছে। এই সম্পর্ক অল্প শিক্ষিত বা অশিক্ষিত লোকের ক্ষেত্রেও প্রযোজ্য। কিন্তু অশিক্ষিত, অল্প শিক্ষিত বা উচ্চ শিক্ষিত লোকের আয় একই জাতীয় নয়। এক্ষেত্রে আয়ের ভিত্তিতে লোকদের মধ্যে একটি ধারাবাহিক পার্থক্য বিদ্যমান থাকা স্বাভাবিক। অনুরূপভাবে বলা যায় যে গাভী দুধ দেয়। দুধের পরিমাণ গাভীর বয়স, খাদ্য ইত্যাদির উপর নির্ভর করে। কিন্তু বয়স এবং খাদ্য এক হলেও সব জাতের গাভী একই পরিমাণ দুধ দেয় না। বাজার থেকে মানুষ খাদ্য দ্রব্য ক্রয় করে। কিন্তু দ্রব্য ক্রয়ের ক্ষেত্রে তার গুণগত এবং পরিমাণগত মান সকল লোকের ক্ষেত্রে এক নয়। একই পদে নিয়োজিত সকল লোকের কর্মদক্ষতা বা ব্যক্তিত্ব একই হয় না। সকল শহরে বসবাসকারি লোকদের আর্থ-সামাজিক চলকের মান একইরূপ হয় না। কাজেই বুঝা যাচ্ছে যে, নমুনা বিন্দুসমূহকে তাদের মধ্যে বিদ্যমান কোনো বৈশিষ্ট্যের ভিত্তিতে একটি অর্থবহ বা ধারাবাহিক পার্থক্য লক্ষ্য করা গেলে গুচ্ছায়ন (clustering) করা যেতে পারে। এরূপ গুচ্ছায়ন করার পদ্ধতি হলো গুচ্ছ বিশ্লেষণ (cluster analysis)।

কোন কোন নমুনা বিন্দু নিয়ে গুচ্ছ করা হবে এ সম্পর্কে কোনো ধরা বাঁধা নিয়ম নেই। তবে কথা হলো যে গুচ্ছের নমুনা বিন্দুসমূহের মধ্যে একটি সামঞ্জস্যতা

(similarity) বজায় থাকা দরকার। এই সামঞ্জস্যতা চিহ্নিত করার নির্দেশকও (criterion) বিভিন্ন গবেষকের ক্ষেত্রে বিভিন্ন হতে পারে। তবে একটি কথা হলো যে গুচ্ছের নমুনা বিন্দুসমূহ চলকের মানের ভিত্তিতে একইরূপ হওয়া উচিত এবং দুই বা ততোধিক গুচ্ছের নমুনা বিন্দুর মধ্যে পার্থক্য বেশি থাকা উচিত। চোখের দৃষ্টিতে একইরূপ নমুনা বিন্দু চিহ্নিত করার ক্ষেত্রে চিত্রের সাহায্য নেয়া যেতে পারে। ধরা যাক n নমুনা বিন্দু (sample objects) এর প্রতিটি হতে p চলকের মান পাওয়া গেছে। এই মানসমূহকে p-মাত্রার (p-dimensional) চিত্রে চিত্রায়িত করা যেতে পারে। এক একটি চলকের মান এক একটি অক্ষে প্রতিস্থাপন করে n বিন্দুর প্রতিটির জন্য চিত্র আঁকা হলে যে বিন্দুগুলো একই এলাকায় পড়বে এবং অন্য এলাকা থেকে বিচ্ছিন্ন থাকবে সেগুলো নিয়ে একটি গুচ্ছ বিবেচনা করা যেতে পারে। গাণিতিকভাবে যে কোনো দুটি বিন্দুর মধ্যে দূরত্ব পরিমাপ করে দূরত্বের সমতার ভিত্তিতে গুচ্ছায়ন করা যেতে পারে। পরবর্তী অনুচ্ছেদসমূহে এ সম্পর্কে আলোচনা করা হবে।

উপরের আলোচনা হতে বুঝা যাচ্ছে যে, যদি $X(n \times p)$ একটি উপাত্ত ম্যাট্রিক্স (data matrix) হয় এবং $X_1, X_2, \cdots, X_n$ যদি n নমুনা বিন্দু হতে প্রাপ্ত p চলকের পরিমাপ নির্দেশ করে তাহলে গুচ্ছ বিশ্লেষণের উদ্দেশ্য হবে $X_1, X_2, \ldots, X_n$-কে n_1 গুচ্ছে ভাগ করা, যেখানে $(n_1 < n)$ n_1 n অপেক্ষা খুব ছোট হবে। এ জাতীয় গুচ্ছ বিশ্লেষণকে এক নমুনাভিত্তিক গুচ্ছ বিশ্লেষণ বলে। আবার গুচ্ছ বিশ্লেষণ m নমুনার ক্ষেত্রেও করা যায়। ধরা যাক X_{ij} $(i = 1, 2, \ldots, n_j \quad j = 1, 2, \cdots, m)$ হলো j-তম নমুনার i-তম একক হতে প্রাপ্ত p চলকের পরিমাপ। এক্ষেত্রে গুচ্ছ বিশ্লেষণের উদ্দেশ্য হলো m নমুনাকে m_1 $(m_1 \leqslant m)$ গুচ্ছে বিভক্ত করা যেন প্রতি গুচ্ছের নমুনা উপাত্তের মধ্যে পার্থক্য কম থাকে কিন্তু এক গুচ্ছের নমুনা উপাত্ত অন্য গুচ্ছের নমুনা উপাত্ত হতে অধিক ভিন্নতর হয়।

উপরের আলোচনা হতে বুঝা যাচ্ছে যে, গুচ্ছ বিশ্লেষণ অন্যান্য উপাত্ত সঙ্কোচন (data reduction) পদ্ধতির মতো একটি। কারণ এই পদ্ধতির মাধ্যমে n নমুনা একক হতে প্রাপ্ত p-চলকের মানভিত্তিক উপাত্তকে n_1 $(n_1 < n)$ গুচ্ছে ভাগ করা হয়। এখানে n_1 এর মান অজানা। এই উপাত্ত সঙ্কোচন পদ্ধতির দ্বারা X ম্যাট্রিক্সের সারির সংখ্যা কমানো হয়ে থাকে। উপাত্ত সঙ্কোচনের অনেক পদ্ধতির মধ্যে একটি হলো প্রধান উপাদান বিশ্লেষণ (principal component analysis)। এই পদ্ধতিতে X ম্যাট্রিক্সের স্তম্ভের সংখ্যা কমানো হয়ে থাকে।

উপরে আলোচিত গুচ্ছ বিশ্লেষণের ব্যবহার বিজ্ঞানের বিভিন্ন অনুচ্ছেদে হয়ে থাকে। জীববিদ্যায় জন্তু এবং উদ্ভিদের শ্রেণিবিন্যাস করার জন্য এর ব্যবহার হয়ে থাকে। সেক্ষেত্রে গুচ্ছ বিশ্লেষণকে বলা হয় নিউমারিকেল ট্যাক্সোনমি

(Numerical taxonomy)। রোগ নিরাময় বিদ্যায় রোগ চিহ্নিতকরণ এবং তার পর্যায় নির্ণয়ের জন্য এই বিশ্লেষণের ব্যবহার হয়ে থাকে। ব্যবসা ক্ষেত্রে লোকজনের ক্রয় অভ্যাস চিহ্নিতকরণের জন্য এবং তার ভিত্তিতে ক্রেতার শ্রেণিবিন্যাস করার জন্য গুচ্ছ বিশ্লেষণ করা হয়। মনস্তত্ত্ববিজ্ঞানে মানুষের ব্যক্তিত্ব পর্যালোচনা করে শ্রেণিবিন্যাস করা হয়। আঞ্চলিক বিশ্লেষণে অঞ্চলের লোকদের আর্থ-সামাজিক চলক পর্যালোচনা করে অঞ্চলের শ্রেণিবিন্যাস করা হয় এই বিশ্লেষণ পদ্ধতির মাধ্যমে।

৭.২ গুচ্ছ বিশ্লেষণের মৌলিক ধাপসমূহ (Basic Steps of Cluster Analysis)

গুচ্ছ বিশ্লেষণ শুরু করার আগে কতগুলো বিষয়ে সিদ্ধান্ত নিতে হয়। আগেই উল্লেখ করা হয়েছে যে কোনো একটি বৈশিষ্ট্যের ভিত্তিতে নমুনা এককসমূহকে গুচ্ছে বিভক্ত করা হলো গুচ্ছ বিশ্লেষণ। সুতরাং কোনো বৈশিষ্ট্যের ভিত্তিতে গুচ্ছ করা হবে তা আগেই চিহ্নিত করা দরকার। আবার গুচ্ছে বিভক্ত করার জন্য চলকের মানভিত্তিক এককসমূহের সাদৃশ্যতা (similarity) বা তাদের মধ্যে দূরত্ব পরিমাপ করার প্রয়োজন হয়। এই দূরত্ব পরিমাপ কিভাবে করা হবে তাও আগেই ঠিক করা প্রয়োজন। তাছাড়া দূরত্ব পরিমাপ করার পরে গুচ্ছায়ন করার জন্য কি নির্দেশক (criterion) ব্যবহার করা হবে সে সম্পর্কেও সিদ্ধান্ত নিতে হয়। এছাড়া গুচ্ছ বিশ্লেষণের শুরুতে কোনো কোনো চলক গুচ্ছ বিশ্লেষণে বিবেচিত হবে তাও নির্দিষ্ট করা প্রয়োজন। কারণ গুরুত্বপূর্ণ চলক বিশ্লেষণ হতে বাদ পড়লে বিশ্লেষিত ফলাফলের গুরুত্ব কমে যাবে বা ঐ বিশ্লেষণ কোনো অর্থবহ তথ্য পরিবেশন করবেনা। যেমন, ব্যবসা সংক্রান্ত গবেষণায় গুচ্ছ বিশ্লেষণ করার ক্ষেত্রে যদি ক্রেতার রুচি এবং আর্থিক সচ্ছলতা বিবেচনা করা না হয়, তাহলে ক্রেতা সাধারণকে গুচ্ছায়ন করা অর্থবহ হবে না।

চলক নির্বাচিত করার পরের প্রশ্ন হলো এককগুলো কিভাবে একই জাতীয় হবে তা সম্পর্কে সিদ্ধান্ত নেয়া। এ পর্যায়ে দুটি এককের মধ্যে পার্থক্য (distance) নির্ণয় বা দুটি এককের মধ্যে ঘনিষ্টতার (closeness) পরিমাপ করার পদ্ধতি চিহ্নিত করা প্রয়োজন। কারণ দুটি এককের মধ্যে পার্থক্য বেশি হলে তারা ভিন্ন ভিন্ন গুচ্ছে অন্তর্ভুক্ত হবে এবং তাদের মধ্যে ঘনিষ্টতা বেশি হলে তারা একই গুচ্ছে অন্তর্ভুক্ত হবে। সুতরাং যে কোনো দুটি এককের মধ্যে দূরত্ব পরিমাপ করা গুচ্ছ বিশ্লেষণে একটি গুরুত্বপূর্ণ কাজ।

এককের সাদৃশ্যতা (similarity) বা দূরত্ব (distance) পরিমাপের বিষয়টিকে প্রধানত দু'ভাগে বিভক্ত করা যায়। এই বিভক্তিকরণ উপাত্তের গুণাগুণের উপর

নির্ভর করে। প্রধান দু'ভাগের একটি হলো : (ক) দূরত্ব জাতীয় পরিমাপ (Distance-type measure এবং অপরটি হলো (খ) অনুরূপ জাতীয় পরিমাপ (Matching-type measure)।

৭.২.১ দূরত্ব-জাতীয় পরিমাপ (Distance-Type Measure) : ধরা যাক n একক্বের প্রতিটি হতে p চলকের মান পরিমাপ করা হয়েছে এবং i-তম (i = 1, 2, ... , n) একক হতে প্রাপ্ত p-চলকের মানসমূহকে $X_i' = (X_{i1}\ X_{i2}\ \dots\ X_{ip})$ ভেক্টর দ্বারা প্রকাশ করা যায়। ধরা যাক d_{ij} ($i \neq i = 1, 2, \dots, p$) হলো দুটি i-তম একক ও j-তম এককের দূরত্ব। যেখানে

$$d_{ij} = \left\{ \sum_{k=1}^{p} | X_{ik} - X_{jk} |^r \right\}^{1/r} \qquad (৭.২.১)$$

এই d_{ij} হলো Minkowski metric এর একটি বিশেষ রূপ। এখানে r = 2 ধরা হলে d_{ij} এর সূত্র দাঁড়ায়

$$d_{ij} = \left\{ \sum_{k=1}^{p} (X_{ik} - X_{jk})^2 \right\}^{1/2} \qquad (৭.২.২)$$

এই d_{ij} হলো i-তম ও j-তম এককের মধ্যে Euclidean দূরত্ব। আবার r = 1 হলে

$$d_{ij} = \sum_{k=1}^{p} | X_{ik} - X_{jk} | \qquad (৭.২.৩)$$

এই শেষোক্ত d_{ij} হলো চিরুবদ্ধিত বা City-block metric।

ধরা যাক নিচের উপাত্তটি হলো পাঁচটি কীটের দৈহিক দৈর্ঘ্য (Lin mm) এবং দৈহিক ওজন (W in gms)

	1	2	3	4	5
L :	35	35	38	35	39
W :	1.3	4.0	3.2	1.0	1.4

উক্ত উপাত্তের ভিত্তিতে Euclidean দূরত্ব নির্ণয়ের সূত্র প্রয়োগ করে নিম্নোক্ত দূরত্ব ম্যাট্রিক্স d পাওয়া যায়।

$$d = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.00 & 2.70 & 3.55 & 0.30 & 4.00 \\ 2.70 & 0.00 & 3.10 & 3.00 & 4.77 \\ 3.55 & 3.10 & 0.00 & 3.72 & 2.06 \\ 0.30 & 3.00 & 3.72 & 0.00 & 4.02 \\ 4.00 & 4.77 & 2.06 & 4.02 & 0.00 \end{bmatrix} \end{array}$$

এখানে লক্ষ্য করলে দেখা যাচ্ছে যে চতুর্থ কীট, দ্বিতীয়, তৃতীয় এবং পঞ্চম কীট অপেক্ষা প্রথম কীটের বেশি ঘনিষ্ট (close)। এখন দেখা যাক দৈর্ঘ্যের পরিমাপ মি. মি-তে না করে সে.মি-তে করা হলে d ম্যাট্রিক্সের মানসমূহ কেমন হয়। অর্থাৎ

	1	2	3	4	5
L :	3.5	3.5	3.8	3.5	3.9
W :	1.3	4.0	3.2	1.0	1.4

হলে Euclidean দূরত্ব ম্যাট্রিক্স হবে

$$d_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.00 & 2.70 & 1.92 & 0.30 & 0.41 \\ 2.70 & 0.00 & 0.85 & 3.00 & 2.63 \\ 1.92 & 0.85 & 0.00 & 2.22 & 1.80 \\ 0.30 & 3.00 & 2.22 & 0.00 & 0.57 \\ 0.41 & 2.63 & 1.80 & 0.57 & 0.00 \end{bmatrix} \end{array}$$

লক্ষ্য করা যাচ্ছে যে d_1 এর ক্ষেত্রেও চতুর্থ কীট প্রথম কীটের বেশি ঘনিষ্ট। এর পরের ঘনিষ্টতা হলো প্রথম ও পঞ্চম কীটের মধ্যে। কিন্তু d হতে লক্ষ্য করা যাচ্ছে যে প্রথম ও পঞ্চম কীটের মধ্যে দূরত্ব সবচেয়ে বেশি। এ থেকে বুঝা যাচ্ছে যে Euclidean দূরত্ব চলকের পরিমাপের একক (unit) দ্বারা প্রভাবিত হয়ে থাকে। অবশ্য এই সমস্যা দূর করার জন্য কেহ কেহ দূরত্ব পরিমাপ করার আগে চলক-গুলোকে আদর্শায়িত (standardised) করার প্রস্তাব করেছেন [Dillon and Goldstein (1984)]।

দূরত্ব নির্ণয় করার অন্য একটি পরিমাপ হলো Mahalanobis D^2। যেখানে

$$D^2=(X_i-X_j)' S^{-1} (X_i-X_j)$$

এখানে X_i ও X_j হলো i-তম ও j-তম একক হতে প্রাপ্ত p চলকের মানের ভেক্টর এবং S হলো আন্ত-গ্রুপ পুলড সহভেদাঙ্ক ম্যাট্রিক্স (Pooled within-group covariance matrix)।

৭.২.২ অনুরূপ জাতীয় পরিমাপ (Matching-Type Measures) : এই পরিমাপ সংশ্লেষণ সহগ নামে অধিক পরিচিত। চলকের মান নমিনাল স্কেলে (nominal scale) পরিমাপ করা হলে, অর্থাৎ গুণগত চলকের ক্ষেত্রে এই সাদৃশ্য পরিমাপ করা হয় এবং এই পরিমাপের মান 0 থেকে 1 হয়। এই পরিমাপ ব্যবহার করার যুক্তিসঙ্গত কারণ হলো যে, চলক গুণগত হওয়ার কারণে দুই বা ততোধিক এককের মানে সাদৃশ্য থাকতে পারে। যেমন, আর্থ-সামাজিক গবেষণার ক্ষেত্রে পারিবারিক উপাত্ত সংগ্রহ করা হলে স্বামী-স্ত্রীর শিক্ষার স্তরে সাদৃশ্য লক্ষ্য করা স্বাভাবিক। কোনো বিশেষ ইস্যুতে মতামত প্রকাশ করতে বলা হলে স্বামী-স্ত্রী একই ইস্যুতে অনুরূপ মতামত প্রকাশ করতে পারে। সেক্ষেত্রে স্বামী-স্ত্রীর মধ্যে সাদৃশ্য বেশি হওয়া স্বাভাবিক।

গুণগত চলকের ক্ষেত্রে সাদৃশ্যতা পর্যালোচনা করার জন্য ধরা যাক কোনো চলকের মান হলো 0 বা 1। ধরা যাক 0 হলে চলকে বৈশিষ্ট্যের অনুপস্থিতির (−) এর নির্দেশক এবং 1 হলো বৈশিষ্ট্যের উপস্থিতির (+) নির্দেশক। দুটি একক হতে প্রাপ্ত এরূপ মানকে 2 × 2 সংযোগ সারণি (contingency table) এর মাধ্যমে প্রকাশ করা যায়। এক্ষেত্রে i-তম এককের মানকে সারিতে এবং j-তম এককের মানকে স্তম্ভে সাজানো হলে সারণির কোষের মানসমূহ হবে উভয় এককের মধ্যে একই বৈশিষ্ট্য বিদ্যমান বা তার বৈপরীত্য নির্দেশকারী সংখ্যা। যেমন, দুইজন লোকের নিম্নলিখিত তথ্যভিত্তিক একটি সংযোগ সারণি বিবেচনা করা যাক।

বৈশিষ্ট্য	নির্দেশক	মান	
		প্রথম একক	দ্বিতীয় একক
শিক্ষা	শিক্ষিত (1)	1	1
	অশিক্ষিত (0)		
পেশা	কৃষি (1)	1	0
	অকৃষি (0)		

ধর্ম	মুসলিম (1) অমুসলিম (0)	1	0
বৈবাহিক অবস্থা	বিবাহিত (1) অবিবাহিত (0)	1	1
বাসস্থান	শহরে (1) গ্রামে (0)	0	1

সারণি ৭.১ ঃ সংযোগ সারণি।

		দ্বিতীয় একক		
প্রথম একক		+	−	মোট
	+	2	2	4
	−	1	0	1
	মোট	3	2	5

উপরিউক্ত সংযোগ সারণি থেকে সংশ্লেষের পরিমাপ বিভিন্নভাবে করা যায়। ঐ পরিমাপগুলো অনুরূপ জাতীয় পরিমাপ হিসেবে ব্যবহৃত হয়। আলোচিত সারণির ক্ষেত্রে এককদ্বয়ের মধ্যে বিদ্যমান (+, +) চিহ্ন ও (−, −) দ্বারা তাদের মধ্যে সাদৃশ্যতা বুঝায়। কাজেই মোট বৈশিষ্ট্যের ভিত্তিতে সদৃশ্যতার সংখ্যাকে মোট বৈশিষ্ট্যের দ্বারা ভাগ করে সংশ্লেষের একটি পরিমাপ পাওয়া যেতে পারে। এখানে এই পরিমাপ হলো $2/5 = 0.40$। আবার অন্যভাবেও এই পরিমাপ পাওয়া যায়। সেক্ষেত্রে এককদ্বয়ের মধ্যে বিদ্যমান (+, +) এর মোট চিহ্ন এবং প্রথম একক বা দ্বিতীয় এককের মধ্যে বিদ্যমান মোট (+)-এর অনুপাত নির্ণয় করা হয়। উক্ত উদাহরণের ক্ষেত্রে এই পরিমাপ হলো $2/(2+2+1) = 0.40$। সদৃশ্যতার আরো অনেক পরিমাপ আছে। নিচে সেগুলো আলোচনা করা হলো।

ধরা যাক সংযোগ সারণিটি হলো নিম্নরূপ :

		দ্বিতীয় একক		
প্রথম একক		+	−	মোট
	+	a	b	a+b
	−	c	d	c+d
	মোট	a+c	b+d	a+b+c+d=n

উপরিউক্ত সারণির ক্ষেত্রে সদৃশ্যতার সহগ (similarity coefficient) হলো

(i) $\frac{a+b}{n}$ (ii) $\frac{a}{a+b+c}$

(iii) $\frac{2a}{2a+b+c}$ (iv) $\frac{2(a+d)}{2(a+b)+b+c}$

(v) $\frac{a}{a+2(b+c)}$ (vi) $\frac{a}{n}$

আলোচিত উদাহরণের ক্ষেত্রে উপরিউক্ত সদৃশ্যতার সহগ নির্ণয়ের সূত্র প্রয়োগ করে পাওয়া যায় (i) 0.80, (ii) 0.40, (iii) 0.57, (iv) 0.36, (v) 0.25 এবং (vi) 0.40 । লক্ষ্য করা যাচ্ছে যে, বিভিন্ন সূত্র হতে সদৃশ্যতার সহগের মান বিভিন্ন হচ্ছে । এর কারণ নির্ভর করে সূত্রে কিভাবে (1) ঋণাত্মক সদৃশ্যতা (0,0) অন্তর্ভুক্ত হয়েছে, (2) চলকগুলোর সাদৃশ্য জোড়াগুলোকে সমান ভর দেয়া হয়েছে কিনা, (3) অসদৃশ্য জোড়াগুলো সদৃশ্য জোড়াগুলোর দ্বিগুণ ভর পেয়েছে কিনা এবং (4) ঋণাত্মক সাদৃশ্যতা (0,0) সূত্র থেকে বাদ দেয়া হয়েছে কিনা, তার উপর । অবশ্য সহগের মান সদৃশ্যতা পর্যালোচনা করার ক্ষেত্রে গুরুত্বপূর্ণ নয় । কারণ চলকের মান (+ , −) এর পরিবর্তে (− , +) ধরা হলেও সহগের মান পরিবর্তিত হতে পারে । যেমন, সারণি ৭.১-এর ক্ষেত্রে (+ , −) এর পরিবর্তে (−, +) ব্যবহার করা হলে নতুন সংযোগ সারণি হয় নিম্নরূপ :

সারণি ৭.২ ঃ সংযোগ সারণি

		দ্বিতীয় একক		
প্রথম একক		+	−	মোট
	+	2	2	4
	−	0	1	1
	মোট	2	3	5

এই সারণির জন্য সদৃশ্যতার সহগগুলো হলো (i) 0.80, (ii) 0.50, (iii) 0.67, (iv) 0.60, (v) 0.33 এবং (vi) 0.40 । দেখা যাচ্ছে যে, সারণি ৭.১ এবং ৭.২ এর ক্ষেত্রে সদৃশ্যতার সহগের মানগুলো এক নয়, যদিও সহগগুলো নির্ণয় করা হয়েছে একই একক হতে । কিন্তু চলকের মান পরিমাপ করার পার্থক্যের কারণে সদৃশ্যতার সহগের মানগুলো ভিন্নতর হয়েছে ।

সদৃশ্যতা পরিমাপ করার জন্য সংশ্লেষাঙ্ক-এর ব্যবহার লক্ষ্য করা যাক । চলক পরিমাণগত (quantitative) হলে সংশ্লেষাঙ্ক নির্ণয় করা হয় । তবে সংশ্লেষাঙ্ক-এর বড় হ[illegible]টি এককের মধ্যে সদৃশ্যতা থাকবে এমন কোনো কথা নয় ।

বিষয়টি একটি উদাহরণের সাহায্যে ব্যাখ্যা করা যাক। ধরা যাক তিনটি একক A, B এবং C হতে চলক x_1, x_2, x_3, x_4, x_5 এবং x_6 এর মান পাওয়া গেছে নিম্নরূপ :

এককসমূহ	চলকের মান					
	x_1	x_2	x_3	x_4	x_5	x_6
A	3	3	8	6	4	6
B	5	5	10	8	6	8
C	9	3	8	6	4	6

উক্ত উপাত্তের ভিত্তিতে A এবং B এর মধ্যে সংশ্লেষাঙ্ক হলো 1। এই সংশ্লেষাঙ্ক হতে বুঝা যায় যে A এবং B এর মধ্যে সংযোগ বেশি। অবশ্য সংযোগ বেশি হলেই যে সদৃশ্যতা থাকবে এমন কোনো কথা নেই। বিষয়টি A এবং C এর মধ্যে সংশ্লেষাঙ্ক নির্ণয় করলেই বুঝা যাবে। এখানে A এবং C এর মধ্যে সংশ্লেষাঙ্ক হলো মাত্র 0.35। অথচ x_1 এর মান ভিন্ন A ও C এর অন্যান্য মান একই। কাজেই এর সংশ্লেষাঙ্ক সংযোগ-এর পরিমাপ হওয়া সত্ত্বেও একে সদৃশ্যতা পর্যালোচনা করার জন্য ব্যবহার করা হয় না।

৭.৩ গুচ্ছ তৈরি (Forming Clusters)

এতক্ষণ সদৃশ্যতা বা দূরত্ব পরিমাপ পদ্ধতি আলোচনা করা হয়েছে। এর পরবর্তী ধাপ হলো কিভাবে গুচ্ছ তৈরি করা হবে। দূরত্ব বা সদৃশ্যতা পরিমাপ করার জন্য যেমন অনেক পদ্ধতি আছে তেমনি গুচ্ছ তৈরি করারও অনেক পদ্ধতি আছে। তবে দুটি পদ্ধতি অধিক ব্যবহৃত হয়। এগুলো হলো (1) Hierarchical পদ্ধতি এবং (2) Partitioning পদ্ধতি। Hierarchical পদ্ধতিকে Hierarchical গুচ্ছ বিশ্লেষণও বলা হয়। এই বিশ্লেষণের জন্য দুটি পদ্ধতি ব্যবহৃত হয়। এগুলো হলো (i) Agglomerative Method এবং (ii) Divisive Method। Hierarchical গুচ্ছায়নের একটি প্রাথমিক বৈশিষ্ট্য হলো যে একটি একক কোনো গুচ্ছে অন্তর্ভুক্ত হলে তাকে ঐ গুচ্ছ হতে সরানো যাবে না এবং অন্য গুচ্ছের কোনো এককের সাথে মিশানো যাবে না। আবার, Divisive গুচ্ছায়নের বৈশিষ্ট্য হলো যে তা n এককসমূহকে এমনভাবে বিভক্ত করে যেন প্রতিটি একক এক একটি গুচ্ছে অন্তর্ভুক্ত হয়। অপরপক্ষে Partitioning পদ্ধতির প্রাথমিক বৈশিষ্ট্য হলো যে কোনো একটি একক একবার এক গুচ্ছে অন্তর্ভুক্ত হলে তা পরে অন্য কোনো গুচ্ছে আবার অন্তর্ভুক্ত হতে পারে। এখানে Hierarchical গুচ্ছায়ন ও Partitioning পদ্ধতি সংক্ষিপ্তভাবে আলোচনা করা হবে। বিস্তারিত জানার জন্য Ball (1971), Cormack (1971), Hartigan (1973, 1975), Jardine (1970),

Gower and Ross (1969), Jardine and Sibson (1971) এবং Everitt (1980) পর্যালোচনা করা যেতে পারে।

৭.৩.১ Agglomerative Method : এই পদ্ধতির শুরুতে বিবেচনা করা হয় যে, প্রতিটি একক ভিন্ন ভিন্ন গুচ্ছে বিভক্ত এবং এককের সংখ্যা যত গুচ্ছের সংখ্যাও তত। তারপর যে কোনো গুচ্ছের সাথে বাকি এককগুলোকে মিশিয়ে বড় গুচ্ছ করা হয়। বড় গুচ্ছ করার প্রক্রিয়া এমনভাবে চলতে থাকে যেন সব এককই একটি গুচ্ছে অন্তর্ভুক্ত হয়। বড় গুচ্ছ করার জন্য প্রথমে দুটি এককে যুক্ত করে একটি একক গুচ্ছ (single cluster) তৈরি করা হয়। পরবর্তী ধাপে এই একক গুচ্ছে তৃতীয় একটি একক অন্তর্ভুক্ত করা হয় বা অন্য দুটি এককের সমন্বয়ে নতুন অন্য একটি গুচ্ছ তৈরি করা হয়। তারপর যে কোনো গুচ্ছে একটি একটি করে একক অন্তর্ভুক্ত করা হয় বা যে কোনো দুটি গুচ্ছকে যুক্ত করা হয়। এই পদ্ধতির বিশেষত্ব হলো যে, একবার গুচ্ছ তৈরি হলে তাকে আর বিভক্ত করা হয় না, প্রয়োজনে ঐ গুচ্ছকে অন্য গুচ্ছের সাথে যুক্ত করা হয়।

এখন প্রশ্ন হলো প্রাথমিক একক গুচ্ছ তৈরির নির্দেশক কি ? বা, পরবর্তী এককসমূহ কি নির্দেশকের ভিত্তিতে গুচ্ছে অন্তর্ভুক্ত হবে ? নির্দেশক (criterion) গুলোর সবই নির্ভর করে জোড়ায় জোড়ায় এককের দূরত্ব বা সদৃশ্যতাসমূহের ম্যাট্রিক্স-এর উপর। এই নির্দেশকগুলোর গুরুত্বপূর্ণ চারটি হলো (1) Single Linkage বা Nearest-Neignbor Method (2) Complete Linkage বা Furthest Neighbor Method, (3) Average Linkage এবং (4) Ward's Error sum of squares Method। এই পদ্ধতিগুলোর মৌলিক পার্থক্য নির্ভর করে সদৃশ্যতা বা দূরত্ব-এর সংজ্ঞায়নের উপর।

Single Linkage বা Nearest Neighbor Method : এই পদ্ধতিতে সবচেয়ে কম দূরত্ববিশিষ্ট দুটি এককে প্রথমে একটি গুচ্ছে অন্তর্ভুক্ত করা হয়। দ্বিতীয় পর্যায়ে তৃতীয় একটি এককে প্রথমোক্ত গুচ্ছে সংযুক্ত করা হয়, বা গুচ্ছে অন্তর্ভুক্ত নয় এমন এককসমূহের মধ্যে যে দুটির দূরত্ব সবচেয়ে কম সে দুটি নিয়ে অন্য একটি গুচ্ছ করা হয়।

এই নতুন গুচ্ছ তৈরি করা বা প্রথম গুচ্ছে একক অন্তর্ভুক্ত করা নির্ভর করে দূরত্বের পরিমাপের উপর। একবার তৈরি করা কোনো গুচ্ছ থেকে ঐ গুচ্ছে অন্তর্ভুক্ত নয় এমন কোনো এককের দূরত্ব সবচেয়ে কম হলে ঐ একক গুচ্ছে অন্তর্ভুক্ত হবে। কিন্তু ইতোমধ্যে তৈরি করা কোনো গুচ্ছ ও অন্য কোনো এককের দূরত্ব অপেক্ষা যদি গুচ্ছে অন্তর্ভুক্ত নয় এমন দুটি এককের দূরত্ব কম হয়, তাহলে শেষোক্ত একক দুটি নিয়ে নতুন গুচ্ছ তৈরি হবে। যতক্ষণ পর্যন্ত সব একক একটি একক গুচ্ছে (single cluster) অন্তর্ভুক্ত না হবে ততক্ষণ পর্যন্ত গুচ্ছায়ন পদ্ধতি চলতে

থাকবে। এই পদ্ধতিতে কোনো পর্যায়ে নতুন গুচ্ছ তৈরি হলে পুরাতন গুচ্ছ ও নতুন গুচ্ছের দূরত্ব সবচেয়ে কম হলে গুচ্ছদ্বয় একটি গুচ্ছে অন্তর্ভুক্ত হবে। একবার কোনো একক একটি গুচ্ছে অন্তর্ভুক্ত হলে তাকে আর গুচ্ছ থেকে বিভক্ত করা যাবে না। সে কারণে এই পদ্ধতিতে গুচ্ছের দূরত্ব হলো নিকটবর্তী এককসমূহের দূরত্বের সমান। এই পদ্ধতিতে গুচ্ছায়ন করার পর গুচ্ছে অন্তর্ভুক্ত এককসমূহকে চিত্রাকারে প্রকাশ করা যায়। ঐ চিত্রকে বলা হয় ডেনড্রোগ্রাম (Dendrogram)।

এখন ৭.২.১ অনুচ্ছেদে আলোচিত Euclidean দূরত্ব ম্যাট্রিক্স d-এর ভিত্তিতে এককসমূহকে গুচ্ছায়ন করে একটি Dendrogram এঁকে দেখানো যেতে পারে। উক্ত উদাহরণের ক্ষেত্রে একক-1 ও একক-4 এর দূরত্ব সবচেয়ে কম (0.30)। সুতরাং, একক-1 ও একক-2 নিয়ে একটি গুচ্ছ হবে। এখন এই গুচ্ছ ও বাকি এককসমূহের দূরত্ব নির্ণয় করা যেতে পারে। d ম্যাট্রিক্স হতে এই দূরত্বসমূহ নিম্নরূপভাবে নির্ণয় করা যায় :

$$d_{(1,\,4)2} = \min\{d_{12},\, d_{42}\} = 2.70$$

$$d_{(1,\,4)3} = \min\{d_{13},\, d_{43}\} = 3.55$$

$$d_{(1,\,4)5} = \min\{d_{15},\, d_{45}\} = 4.00$$

উপরিউক্ত দূরত্বসমূহ হলো গুচ্ছের সাথে অন্যান্য এককের দূরত্ব। এগুলোকে বলা হয় গুচ্ছ-এককের দূরত্ব (cluster-unit distance)। এখন গুচ্ছ-এককের দূরত্ব ও আন্তঃএকক দূরত্বের ভিত্তিতে নতুন দূরত্ব ম্যাট্রিক্স, ধরা যাক d_2 পাওয়া যায়। এই d_2 হলো

$$d_2 = \begin{array}{c} \\ 1,4 \\ 2 \\ 3 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 1,4 & 2 & 3 & 5 \end{array} \\ \begin{bmatrix} 0.00 & 2.70 & 3.55 & 4.00 \\ 2.70 & 0.00 & 3.10 & 4.77 \\ 3.55 & 3.10 & 0.00 & 2.06 \\ 4.00 & 4.77 & 2.06 & 0.00 \end{bmatrix} \end{array}$$

এই d_2 ম্যাট্রিক্স-এর সবচেয়ে ছোট মান হলো একক-3 ও একক-5 এর দূরত্ব $d_{35} = 2.06$। সুতরাং অন্য কোনো একক আগের গুচ্ছে অন্তর্ভুক্ত না হয়ে একক-3 ও একক-5 নিয়ে নতুন একটি গুচ্ছ তৈরি হবে।

এ পর্যায়ে প্রথম গুচ্ছের সাথে বাকি এককসমূহের দূরত্ব (একক-2), দ্বিতীয় গুচ্ছের সাথে বাকি এককসমূহের দূরত্ব (একক-2) এবং প্রথম গুচ্ছ ও দ্বিতীয় গুচ্ছের দূরত্ব নির্ণয় করতে হবে এবং যেগুলোর দূরত্ব সবচেয়ে কম সেগুলো গুচ্ছভুক্ত হবে। এখানে এই দূরত্বগুলো হলো :

$$d_{(14)2} = \min\{d_{12}, d_{42}\} = 2.70$$

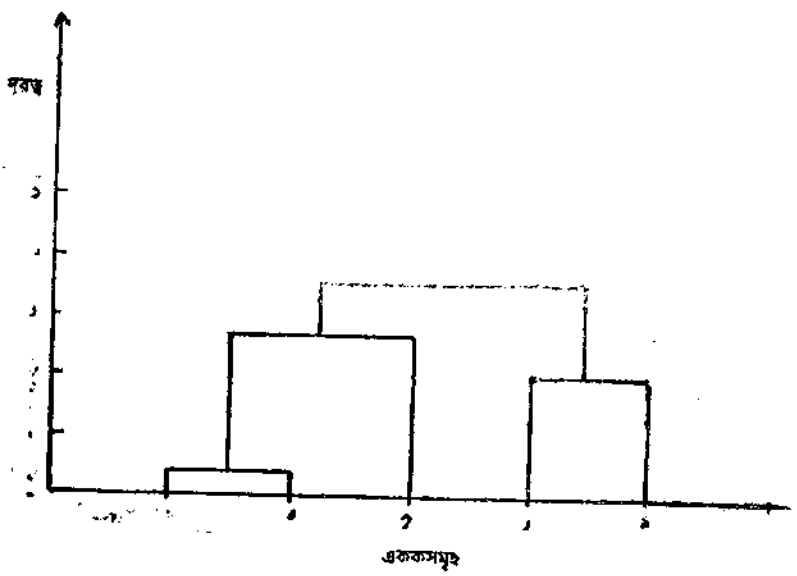
$$d_{(3,5)2} = \min\{d_{23}, d_{25}\} = 3.10$$

$$d_{(14)(35)} = \min\{d_{13}, d_{34}, d_{15}, d_{45}\} = 3.55$$

এই গুচ্ছ-গুচ্ছ দূরত্ব ও গুচ্ছ-একক দূরত্ব নিয়ে নতুন দূরত্ব ম্যাট্রিক্স, ধরা যাক d_3, পাওয়া যায় নিম্নরূপ :

$$d_3 = \begin{array}{c} \\ 14 \\ 2 \\ 35 \end{array} \begin{array}{c} \begin{array}{ccc} 14 & 2 & 35 \end{array} \\ \begin{bmatrix} 0.00 & 2.70 & 3.55 \\ 2.70 & 0.00 & 3.10 \\ 3.55 & 3.10 & 0.00 \end{bmatrix} \end{array}$$

লক্ষ্য করা যাচ্ছে যে, $d_{(14)2} = 2.70$ হলো সবচেয়ে ছোট। কাজেই একক-2 কে প্রথম গুচ্ছে অন্তর্ভুক্ত করতে হবে এবং সবশেষে দ্বিতীয় গুচ্ছকে প্রথম গুচ্ছে অন্তর্ভুক্ত করে সব কয়টি এককের জন্য একটি গুচ্ছ পাওয়া যাবে। নিচে এককসমূহের গুচ্ছভুক্তি Dendrogram-এর সাহায্যে দেখানো হলো এবং সকল একক এক গুচ্ছভুক্ত হওয়ায় দূরত্ব হবে 3.10।



চিত্র ৭.১ : Single Linkage Dendrogram.

Complete Linkage বা Furthest Neighbor Method : এই পদ্ধতি সম্পূর্ণরূপে Single Linkage পদ্ধতির বিপরীত। তবে এক্ষেত্রে প্রথম গুচ্ছ তৈরি হয়

Single Linkage পদ্ধতির ন্যায়। পরবর্তী পর্যায়গুলোতে একক বা গুচ্ছ ইতোমধ্যে তৈরি করা গুচ্ছে অন্তর্ভুক্ত হওয়ার ক্ষেত্রে একক জোড়ার বা একক গুচ্ছের দূরত্ব নির্ণয় করার সময় সবচেয়ে বেশি দূরত্ব বিবেচনা করা হয়। যেমন, Single Linkage পদ্ধতিতে আলোচিত উদাহরণের ক্ষেত্রে একক-1 ও একক-4 এর দূরত্ব কম হওয়াতে ঐ দুটি একক প্রথম পর্যায়ে গুচ্ছভুক্ত হলো। তারপর ঐ গুচ্ছের সাথে অন্য এককসমূহের দূরত্ব নির্ণয় করা হবে নিম্নরূপভাবে :

$$d_{(14)2} = \max\{d_{12}, d_{24}\} = 3.00$$

$$d_{(14)3} = \max\{d_{13}, d_{34}\} = 3.72$$

$$d_{(14)5} = \max\{d_{15}, d_{45}\} = 4.00$$

সুতরাং, d_2 ম্যাট্রিক্স হবে

$$d_2 = \begin{array}{c} \\ 14 \\ 2 \\ 3 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 14 & 2 & 3 & 5 \end{array} \\ \begin{bmatrix} 0.00 & 3.00 & 3.72 & 4.00 \\ 3.00 & 0.00 & 3.10 & 4.77 \\ 3.72 & 3.10 & 0.00 & 2.06 \\ 4.00 & 4.77 & 2.06 & 0.00 \end{bmatrix} \end{array}$$

এখন একক-3 ও একক-5 এর দূরত্ব কম হওয়াতে তারা দ্বিতীয় গুচ্ছভুক্ত হবে। তারপর প্রাপ্ত দুই গুচ্ছ হতে বাকি এককের এবং এই দুই গুচ্ছের দূরত্ব নির্ণয় করতে হবে। দূরত্ব নির্ণয় করার ক্ষেত্রে উপরিউক্ত নিয়ম অনুসরণ করতে হয়। যেমন,

$$d_{(14)2} = 3.00$$

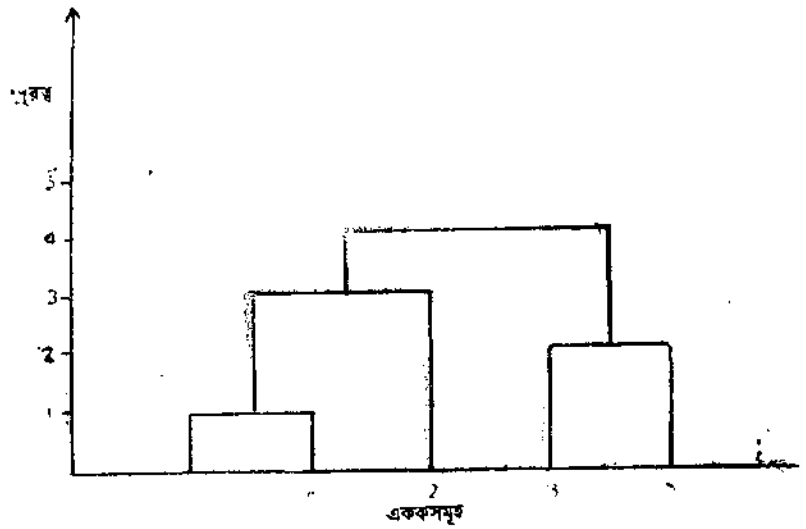
$$d_{(35)2} = \max\{d_{23}, d_{25}\} = 4.77$$

$$d_{(14)(35)} = \max\{d_{13}, d_{15}, d_{34}, d_{45}\} = 4.00$$

সুতরাং d_3 ম্যাট্রিক্স হবে

$$d_3 = \begin{array}{c} \\ 14 \\ 2 \\ 35 \end{array} \begin{array}{c} \begin{array}{ccc} 14 & 2 & 35 \end{array} \\ \begin{bmatrix} 0.00 & 3.00 & 4.00 \\ 3.00 & 0.00 & 4.77 \\ 4.00 & 4.77 & 0.00 \end{bmatrix} \end{array}$$

এখন $d_{(14)2} = 300$ সবচেয়ে কম হওয়ায় একক-2 প্রথম তৈরি করা গুচ্ছে অন্তর্ভুক্ত হবে। এক্ষেত্রে Dendrogram হবে নিম্নরূপ :



চিত্র ৭.২ : Complete Linkage Dendogram.

এই dendogram এবং single linkage পদ্ধতিতে প্রাপ্ত dendogram-এর আকার একইরূপ। পার্থক্য হলো একক-একক দূরত্ব বা একক-গুচ্ছ দূরত্ব-এর পরিমাপ ভিন্নতর। এই Dendogram-এর সদৃশ্যতা সম্বন্ধে আরো জানার জন্য Sneath (1957) Johnson (1967) আলোচনা করা যেতে পারে।

Average Linkage : এই পদ্ধতিতেও আগের দুই পদ্ধতির ন্যায় গুচ্ছায়ন করা হয়। কিন্তু গুচ্ছ-গুচ্ছ বা গুচ্ছ-একক দূরত্ব নির্ণয় করার ক্ষেত্রে তাদের ন্যূনতম (single linkage method) বা সর্বোচ্চ (complete linkage method) দূরত্ব বিবেচনা না করে দূরত্বসমূহের গড় নির্ণয় করা হয়। গড় নির্ণয় করার ক্ষেত্রে গাণিতিক গড় ছাড়াও অন্যান্য পদ্ধতি ব্যবহার করা যায়। এরূপ একটি গড় নির্ণয়ের সূত্র হলো

$$\frac{1}{n_i n_j} \sum_i \sum_j d_{ij}$$

এখানে d_{ij} হলো একটি গুচ্ছের i-তম একক এবং অন্য গুচ্ছের j-তম এককের দূরত্ব, n_i ও n_j হলো গুচ্ছদ্বয়ে অন্তর্ভুক্ত এককের সংখ্যা, দুটি গুচ্ছে অন্তর্ভুক্ত এককসমূহ নিয়ে যত জোড়া করা যায় সে সব জোড়ার দূরত্বের যোগফল নিতে হয়।

উদাহরণ হিসেবে ৭.২.১ অনুচ্ছেদে আলোচিত Euclidean দূরত্ব ম্যাট্রিক্স d-এর ভিত্তিতে Average linkage-এর মাধ্যমে গুচ্ছায়ন করার জন্য দূরত্ব নির্ণয় করতে পারি। আগেই লক্ষ্য করা গেছে যে, একক-1 ও একক-4 নিয়ে প্রাথমিক গুচ্ছ তৈরি হয়। সুতরাং ঐ প্রাথমিক গুচ্ছ হতে অন্যান্য এককের দূরত্ব হবে নিম্নরূপ :

$$d_{(14)2} = \frac{1}{2\times 1}[d_{12} + d_{24}] = 2.85$$

$$d_{(14)3} = \frac{1}{2\times 1}[d_{13} + d_{34}] = 3.64$$

$$d_{(14)5} = \frac{1}{2\times 1}[d_{15} + d_{45}] = 4.01$$

সুতরাং d_2 ম্যাট্রিক্স হবে

$$d_2 = \begin{array}{c} \\ 14 \\ 2 \\ 3 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 14 & 2 & 3 & 5 \end{array} \\ \begin{bmatrix} 0.00 & 2.85 & 3.64 & 4.00 \\ 2.85 & 0.00 & 3.10 & 4.77 \\ 3.64 & 3.10 & 0.00 & 2.06 \\ 4.00 & 4.77 & 2.06 & 0.00 \end{bmatrix} \end{array}$$

এই d_2 ম্যাট্রিক্স-এর সবচেয়ে ছোট মান একক-3 এবং একক-5 এর দূরত্ব হওয়াতে উক্ত দুটি একক নিয়ে নতুন গুচ্ছ তৈরি হবে। এখন প্রথম গুচ্ছ, দ্বিতীয় গুচ্ছ ও একক-2 এর মধ্যে দূরত্ব হবে নিম্নরূপ :

$$d_{(35)2} = \frac{1}{2\times 2}[d_{23} + d_{25}] = 3.94$$

$$d_{(14)2} = 2.85$$

$$d_{(14)35} = \frac{1}{2\times 2}[d_{13} + d_{15} + d_{34} + d_{45}] = 3.82$$

এই পর্যায়ে d_3 ম্যাট্রিক্স হলো

$$d_3 = \begin{array}{c} \\ 14 \\ 2 \\ 35 \end{array} \begin{array}{c} \begin{array}{ccc} 14 & 2 & 35 \end{array} \\ \begin{bmatrix} 0.00 & 2.85 & 3.82 \\ 2.85 & 0.00 & 3.94 \\ 3.82 & 3.94 & 0.00 \end{bmatrix} \end{array}$$

উক্ত d_3 ম্যাট্রিক্স-এর মান হতে বলা যায় একক-2 প্রথম গুচ্ছের অন্তর্ভুক্ত হবে। কাজেই উক্ত গুচ্ছায়নের ক্ষেত্রেও চিত্র ৭.১-এর ন্যায় **dendogram** হবে।

এই পদ্ধতিতে গুচ্ছায়নের পর ফলাফলকে সারণি ৭.৩-এ উপস্থাপন করে দেখানো যেতে পারে।

সারণি ৭.৩ ঃ Average linkage পদ্ধতিতে গুচ্ছায়নের ফলাফল।

ধাপ	গুচ্ছভুক্তি		দূরত্ব	কোন ধাপে	প্রথম গুচ্ছভুক্তি	পরবর্তী
	প্রথম একক	দ্বিতীয় একক		প্রথম একক	দ্বিতীয় একক	ধাপ
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	1	4	0.30	0	0	3
2	3	5	2.06	0	0	4
3	1	2	2.85	1	0	4
4	1	3	3.94	3	2	0

এখানে প্রথম লাইনে গুচ্ছায়নের প্রথম ধাপের ফলাফল লক্ষ্য করা যাচ্ছে। প্রথম ধাপে একক-1 ও একক-4 গুচ্ছভুক্ত হলো। এই দুই এককের দূরত্ব 0.30। এটি দূরত্ব নির্দেশিত স্তম্ভে দেখানো হয়েছে। পরবর্তী কোন ধাপে এই গুচ্ছের সাথে অন্য একক সংযুক্ত হয়েছে তা শেষ স্তম্ভে দেখানো হয়েছে। লক্ষ্য করা যাচ্ছে যে, তৃতীয় ধাপে প্রথম গুচ্ছের সাথে একক-2 সংযুক্ত হয়েছে। আরো লক্ষ্য করা যাচ্ছে যে, একক-1 ও একক-4 প্রাথমিক পর্যায়ে (0 ধাপে) গুচ্ছভুক্ত হয়েছে। আবার একক-3 ও একক-5 প্রাথমিক পর্যায়ে (0 ধাপে) গুচ্ছভুক্ত হয়েছে। কোন একক কোন ধাপে প্রথম গুচ্ছভুক্ত হয় তা কোন ধাপে প্রথম গুচ্ছভুক্তি স্তম্ভদ্বয় দ্বারা বুঝানো হয়েছে। এখানে তৃতীয় ধাপের (5) স্তম্ভের মান 1 দ্বারা বুঝানো হয়েছে যে, একক-1 প্রাথমিক পর্যায়ে প্রথম গুচ্ছভুক্ত হয়। আবার দ্বিতীয় ধাপের শেষ স্তম্ভের প্রতি লক্ষ্য করলে বুঝা যায় যে, একক-3 ও একক-5 চতুর্থ ধাপে আবার গুচ্ছভুক্ত হয়েছে।

গুচ্ছভুক্তির ফলাফলকে আড়াআড়ি I cicle চিত্র (Vertical icicle plot) দ্বারা চিত্র ৭.৩ উপস্থাপন করা যায়। প্রথমে এককসমূহকে ভিন্ন ভিন্ন গুচ্ছে বিবেচনা করে দেখানো যায়। যেমন

	একক	একক	একক	একক	একক
ধাপ	5	3	2	4	1
1+	× × × × × × × × × × × × × × × × × × × ×				
2+	× × × × × ×		× × × × × × × × × × × ×		
3+	× × × × × ×		×	× × × × × ×	
4+	×	×	×	× × × × × ×	

চিত্র ৭.৩ : Icicle চিত্র।

এখানে পাঁচটি একক পাঁচ গুচ্ছে বিভক্ত। গুচ্ছায়নের প্রথম ধাপে সবচেয়ে নিকটবর্তী দুটি একক (এখানে একক-1 ও একক-4) গুচ্ছভুক্ত হয়েছে। এটি চতুর্থ ধাপে একক-1 ও একক-4 এর মধ্যে অবিচ্ছিন্ন চিহ্ন দ্বারা দেখানো হয়েছে। এর পরবর্তী ধাপে একক-3 ও একক-5 গুচ্ছভুক্ত হয়েছে। তারপরে একক-2 প্রথম গুচ্ছের অন্তর্ভুক্ত হয়েছে। সবশেষে সকল একক একটি গুচ্ছভুক্ত হয়েছে।

উপরে icicle চিত্র আড়াআড়িভাবে আঁকা হয়েছে। অনেক সময় এককের সংখ্যা বেশি হলে প্রতিটি এককে ভিন্ন ভিন্ন গুচ্ছে দেখানোর জন্য আড়াআড়িভাবে চিত্র আঁকা অসুবিধাজনক। সেক্ষেত্রে icicle চিত্র খাড়া (horizontal) করে আঁকা যায়। অবশ্য উভয় ক্ষেত্রেই চিত্র আঁকার একই পদ্ধতি অনুসরণ করা হয়। নিচে horizontal icicle চিত্র এঁকে দেখানো হলো :

		গুচ্ছ সংখ্যা				
		1	2	3	4	5
একক	ধাপ	+	+	+	+	+
একক-5		× × × × × × × ×				
		× × × × × ×				
একক-3		× × × × × × × ×				
		×				
একক-2		× × × × × × × ×				
		× × ×				
একক-4		× × × × × × ×				
		× × × × × × ×				
একক-1		× × × × × × ×				

চিত্র ৭.৪ : Horizontal icicle।

উপরে আলোচিত Average linkage পদ্ধতিকে আবার দুটি ভাগে বিভক্ত করা যায়। (১) Average linkage between group পদ্ধতি, (২) Average linkage within group পদ্ধতি। প্রথমোক্ত পদ্ধতির ক্ষেত্রে, ধরা যাক একক-1 ও একক-4 একটি গুচ্ছে অন্তর্ভুক্ত এবং একক-3 ও একক-5 একটি গুচ্ছে অন্তর্ভুক্ত। এখন উক্ত দুটি গুচ্ছের দূরত্ব নির্ণয় করার ক্ষেত্রে দুই গুচ্ছের দুটি এককের যত জোড়া করা যায় সে সব জোড়ার দূরত্বের গড় নিতে হয়। অর্থাৎ

$$d_{(14)35} = \frac{1}{4}\left[d_{13} + d_{15} + d_{34} + d_{45}\right]$$

কাজেই এই পদ্ধতি single linkage বা complete linkage পদ্ধতি হতে ভিন্নতর। এই পদ্ধতিতে সকল জোড়ার দূরত্বের ভিত্তিতে গুচ্ছের দূরত্ব নির্ণয় করা হয় বলে এটি অন্য দুই পদ্ধতি অপেক্ষা অধিক গ্রহণযোগ্য।

Average linkage within group পদ্ধতির ক্ষেত্রে গুচ্ছের দূরত্ব নির্ণয় করার সময় দুই গুচ্ছের এককসমূহের মধ্যে সকল জোড়ার দূরত্বের গড় নির্ণয় করতে হয়। যেমন,

$$d_{(14)(35)} = \frac{1}{6}\left[d_{14} + d_{13} + d_{15} + d_{34} + d_{35} + d_{45}\right]$$

Ward's error sum of squares method : এই পদ্ধতির প্রস্তাব করেছেন Ward (1963)। তাঁর মতে গুচ্ছে অন্তর্ভুক্ত এককসমূহের সব কয়টি চলকের জন্য চলকের গড় থেকে নির্ণয় করা ব্যবধানের বর্গের মোট (Total sum of squared deviation from the mean of the cluster) নির্ণয় করতে হয়। গুচ্ছ করার সময় যে গুচ্ছদ্বয়ের ক্ষেত্রে ব্যবধানের বর্গের মোট কম হবে ঐ গুচ্ছদ্বয় দ্বারা একটি নতুন গুচ্ছ তৈরি হবে।

এই পদ্ধতিকে অন্যভাবেও প্রকাশ করা যায়। যেমন, প্রতি গুচ্ছের চলকসমূহের গড় নির্ণয় করে ঐ গড়ের ভিত্তিতে প্রতি চলকের জন্য Euclidean দূরত্বের বর্গ নির্ণয় করতে হয়। তারপর সব চলকের জন্য দূরত্ব যোগ করে নিতে হয়। যে এককসমূহের জন্য এই মোট দূরত্ব কম হবে ঐগুলো একটি গুচ্ছে অন্তর্ভুক্ত হবে।

আলোচিত পদ্ধতি ৭.২.১ অনুচ্ছেদে উল্লেখিত উদাহরণের ক্ষেত্রে প্রয়োগ করা যাক। এখানে

চলক	এককসমূহ				
	1	2	3	4	5
L :	35	35	38	35	39
W :	1.3	4.0	3.2	1.0	1.4

নিয়ম অনুযায়ী প্রথমে পাঁচটি এককে পাঁচটি গুচ্ছে বিবেচনা করতে হয়। ঐক্ষেত্রে আন্তঃগুচ্ছ (within cluster) দূরত্ব হলো শূন্য। এখন এককসমূহকে জোড়ায় জোড়ায় গুচ্ছভুক্ত বিবেচনা করা যাক। নিচে ৭.৪ সারণিতে জোড়ায় জোড়ায়

সারণি ৭.৪ ঃ Ward's পদ্ধতিতে গুচ্ছায়নের প্রাথমিক ধাপ।

দুই এককের গুচ্ছ	চলকের গড়		ব্যবধানের বর্গের মোট
	L	W	(সকল চলকের জন্য)
12	35	2.65	3.645
13	36.5	2.25	6.305
14	35	1.15	0.045
15	37	1.35	8.005
23	36.5	3.60	4.820
24	35	2.25	4.500
25	37	2.70	11.380
34	36.5	2.10	6.920
35	38.5	2.30	2.120
45	37.0	1.20	8.080

গুচ্ছভুক্ত এককসমূহ, চলকের গুচ্ছ গড় এবং গুচ্ছ গড়ের ভিত্তিতে ব্যবধানের বর্গের মোট দেখানো হলো ঃ লক্ষ্য করা যাচ্ছে যে, একক-1 ও একক-4 এর ক্ষেত্রে ব্যবধানের বর্গের মোট সবচেয়ে কম। সুতরাং 1 ও 4 এককদ্বয়কে একটি গুচ্ছভুক্ত করা যায়। এখন এই গুচ্ছের সাথে বাকি তিনটি একক (2, 3 এবং 5) সংযুক্ত করা হলে ব্যবধানের বর্গের মোট হবে নিম্নরূপ, সারণি ৭.৫। দেখা যাচ্ছে যে ব্যবধানের বর্গের মোট একক-3 ও একক-5 এর ক্ষেত্রে সবচেয়ে কম। সুতরাং এই ধাপে কোনো এককে প্রথম গুচ্ছে অন্তর্ভুক্ত না করে 3 ও 5 এককদ্বয়কে একটি নতুন গুচ্ছভুক্ত করতে হবে।

সারণি ৭.৫ ঃ Ward's পদ্ধতিতে গুচ্ছায়নের দ্বিতীয় ধাপ।

গুচ্ছের এককসমূহ	চলকের গড়		ব্যবধানের বর্গের মোট (সকল চলকের জন্য)
	L	W	
142	35	2.10	5.460
143	36	1.83	8.847
145	36.3	1.23	10.754
23	36.5	3.60	4.820
25	37	2.70	11.380
35	38.5	2.30	2.120

এ পর্যায়ে লক্ষ্য করতে হবে বাকি এক্ক-2 প্রথম গুচ্ছভুক্ত হবে না কি দ্বিতীয় গুচ্ছভুক্ত হবে। অথবা লক্ষ্য করতে হবে প্রথম ও দ্বিতীয় গুচ্ছ একটি গুচ্ছভুক্ত হয় কিনা। এতদুদ্দেশ্যে ব্যবধানের বর্গের মোট ৭.৬ সারণিতে দেখানো হলো।

সারণি ৭.৬ ঃ Ward's পদ্ধতিতে গুচ্ছায়নের তৃতীয় ধাপ।

গুচ্ছের এককসমূহ	চলকের গড়		ব্যবধানের বর্গের মোট (সকল চলকের জন্য)
	L	W	
1435	36.8	1.73	15.738
352	37.3	2.87	12.214
142	35	2.10	5.460

দেখা যাচ্ছে যে 1, 4 এবং 2 এককত্রয়ের ক্ষেত্রে ব্যবধানের বর্গের মোট কম। সুতরাং এই পর্যায়ে এক্ক-2 প্রথম গুচ্ছে অন্তর্ভুক্ত হবে। এখন সকল এক্ক এক গুচ্ছভুক্ত হলে ব্যবধানের বর্গের মোট হবে 22.328।

এই গুচ্ছায়নের ক্ষেত্রেও Dendogram চিত্র ৭.১-এর অনুরূপ হবে। তবে, এক্ষেত্রে X-অক্ষে এককসমূহ এবং Y-অক্ষে ব্যবধানের বর্গের মোট চিত্রায়িত করতে হয়। আলোচিত উপাত্তের ক্ষেত্রে গুচ্ছায়ন একইরূপ হয়েছে। বিভিন্ন পদ্ধতিতে গুচ্ছের দূরত্ব পরিমাপ করা হয়েছে ভিন্ন ভিন্ন পদ্ধতিতে। আলোচিত পদ্ধতিসমূহের একটি তুলনামূলক চিত্র ৭.৭ সারণিতে দেয়া হলো।

সারণি ৭.৭ ঃ গুচ্ছায়ন পদ্ধতিসমূহের তুলনামূলক ফলাফল।

ধাপ	Single Linkage				Complete Linkage			
	গুচ্ছ সংখ্যা	দূরত্ব	এককের সংখ্যা	একক	গুচ্ছ সংখ্যা	দূরত্ব	এককের সংখ্যা	একক
1	4	0.30	2	1, 4	4	0.30	2	1, 4
2	3	2.06	2	3, 5	3	2.06	2	3, 5
3	2	2.70	3	1, 2, 4	2	3.00	3	1,4,2
4	1	3.10	5	All	1	4.77	5	All
	Average Linkage				Ward's			
1	4	0.30	2	1, 4	4	0.045	2	1, 4
2	3	2.06	2	3, 5	3	2.120	2	3, 5
3	2	2.85	3	1, 2, 4	2	5.460	3	1,4,2
4	1	3.89	5	All	1	22.328	5	All

উপরে আলোচিত Ward's পদ্ধতি প্রায়ই ব্যবহৃত হয়ে থাকে। এই পদ্ধতি ছাড়া আরো দুটি পদ্ধতি, যেমন (1) Centroid পদ্ধতি এবং (2) Median পদ্ধতি গুচ্ছায়নের জন্য ব্যবহৃত হয়। এই দুই পদ্ধতির ক্ষেত্রে Squared Euclidean দূরত্ব ব্যবহার করা হয়। Centroid পদ্ধতিতে দুটি গুচ্ছের দূরত্ব হলো গুচ্ছদ্বয়ে অন্তর্ভুক্ত চলকসমূহের গড়ের দূরত্ব। এই পদ্ধতির একটি অসুবিধা হলো যে, কোনো এক দূরত্বের ভিত্তিতে গুচ্ছায়ন করা হলে পরবর্তী ধাপে গুচ্ছের দূরত্ব কমে যেতে পারে। কারণ, পরবর্তী ধাপ অপেক্ষা পূর্ববর্তী ধাপে অধিক সাদৃশ্য একক গুচ্ছভুক্ত হয়।

Centroid পদ্ধতির ক্ষেত্রে একটি অন্তর্ভুক্ত গুচ্ছের centroid হলো দুটি ভিন্ন ভিন্ন গুচ্ছের ভরারোপিত সংযোগ। এক্ষেত্রে ভর গুচ্ছের আকারের সমানুপাতিক। Median পদ্ধতিও centtoid পদ্ধতির ন্যায়। তবে অন্তর্ভুক্ত কোনো গুচ্ছের centroid নির্ণয় করার সময় দুটি গুচ্ছকে সমান ভরের মাধ্যমে ভরারোপিত করে সংযোগ করা হয়।

৭.৩.২ Divisive Method : এই পদ্ধতিতে প্রথমে সকল এককে দুটি ভাগে বিভক্ত করা হয়। বিভক্তিকরণ সম্ভব হলে এককসমূহ এক গুচ্ছ হতে অন্য গুচ্ছে

স্থানান্তরিত হয় অথবা গুচ্ছভুক্ত এককসমূহ বিভক্ত হয়ে আরো অধিক ভাল উপগুচ্ছে বিভক্ত হয়। এই পদ্ধতির বিশেষত্ব হলো একবার কোনো একক একটি গুচ্ছভুক্ত হলে তা আবার ঐ গুচ্ছ হতে বিভক্ত হয়ে নতুন করে অন্য গুচ্ছে অন্তর্ভুক্ত হতে পারে। তবে এই পদ্ধতির মূল সমস্যা হলো সকল এককে কিভাবে দুটি ভাগে বিভক্ত করা হবে তা নির্ধারণ করা। n আকারের এককের ক্ষেত্রে দুটি দুটি করে উপগুচ্ছে বিভক্ত করা যায় $(2^{n-1}-1)$ ভাবে। এ কারণে এই পদ্ধতির ক্ষেত্রে এই উপগুচ্ছে বিভক্ত করার ব্যাপারে একটি সাবধানতা অবলম্বন করতে হয়। একবার উপগুচ্ছ করা সম্ভব হলে গুচ্ছ-বিভক্তিকরণ অসুবিধাজনক নয়। এখানে প্রাথমিক উপগুচ্ছ তৈরি এবং তৈরি গুচ্ছকে কিভাবে বিভক্ত করা যায় তা আলোচনা করা হবে।

MaC Naughton-Smith et al. (1962) একটি পদ্ধতি আলোচনা করেছেন। এই পদ্ধতি ৭.২.১ অনুচ্ছেদে আলোচিত d ম্যাট্রিক্স-এর ভিত্তিতে ব্যাখ্যা করা যাবে। এই পদ্ধতিকে বলা হয় Splinter-Average Distance Method। এই পদ্ধতিতে প্রথমে অন্য সব একক হতে যে কোনো একটি এককের গড় দূরত্ব নির্ণয় করা হয়। যে এককের গড় দূরত্ব অন্য সব এককের গড় দূরত্ব অপেক্ষা বেশি তাকে বিভক্ত করা হয় এবং এই একক দ্বারা Splinter গ্রুপ তৈরি হয়। যেমন, d ম্যাট্রিক্স-এর ক্ষেত্রে

	A	B	C	D	E
d = A	0.00	2.70	3.55	0.30	4.00
B	2.70	0.00	3.10	3.00	4.77
C	3.55	3.10	0.00	3.72	2.06
D	0.30	3.00	3.72	0.00	4.02
E	4.00	4.77	2.06	4.02	0.00

অন্যান্য একক হতে E-এর গড় দূরত্ব 3.7125। এই গড় দূরত্ব অন্যান্য এককের গড় দূরত্ব অপেক্ষা বেশি বলে {E} একটি গ্রুপে বিবেচিত হবে এবং {A, B, C, D} অন্য একটি গ্রুপে বিবেচিত হবে। এখানে {E}-কে বলা হয় Splinter গ্রুপ। নিচে Splinter গ্রুপ হতে অন্যান্য এককের গড় দূরত্ব এবং মূল গ্রুপের গড় দূরত্ব দেখানো হলো :

একক	Splinter গ্রুপ হতে গড় দূরত্ব	মূল গ্রুপের গড় দূরত্ব	উভয় দূরত্বের পার্থক্য
A	4.00	2.1833	−1.8167
B	4.77	2.9333	−1.8367
C	2.06	3.4567	1.3967
D	4.02	2.3400	−1.6800

দেখা যাচ্ছে যে C-এর ক্ষেত্রে মূল গ্রুপের গড় দূরত্ব অপেক্ষা Splinter গ্রুপ হতে গড় দূরত্ব কম। এ কারণে C-কে মূল গ্রুপ হতে বিভক্ত করে Splinter গ্রুপে সংযুক্ত করতে হবে। এ পর্যায়ে Splinter গ্রুপে একক হবে {C, E} এবং মূল গ্রুপের একক হলো {A, B, D}। এখন Splinter গ্রুপ হতে অন্যান্য এককের গড় দূরত্ব এবং Splinter গ্রুপের এককদ্বয়ের গড় দূরত্ব নির্ণয় করে নিচে দেখানো হলো।

একক	Splinter গ্রুপ হতে গড় দূরত্ব	মূল গ্রুপের গড় দূরত্ব	উভয় দূরত্বের পার্থক্য
A	3.755	1.50	−2.255
B	3.935	2.85	−1.085
D	3.870	1.65	−2.220

দেখা যাচ্ছে যে, সকল এককের [A, B, D] ক্ষেত্রে মূল গ্রুপের গড় দূরত্ব অপেক্ষা Splinter গ্রুপের গড় দূরত্ব কম। সুতরাং এ অবস্থায় আর কোনো একক Splinter গ্রুপে যুক্ত হবে না।

উপরিউক্ত পদ্ধতিতে এককসমূহকে দুটি গুচ্ছে বিভক্ত করা হয়েছে। কিন্তু এখানেই গুচ্ছায়ন কাজ শেষ হয় নি। পুনরায় দুই গুচ্ছের এককসমূহকে একই পদ্ধতিতে বিভক্ত করতে হবে। এখন {A, B, D} এককের গুচ্ছকে বিভক্ত করা যাক। এক্ষেত্রে A, B ও D এর জন্য দূরত্ব ম্যাট্রিক্স হবে

$$
\begin{array}{c} \\ d = \mathrm{A} \\ \mathrm{B} \\ \mathrm{D} \end{array}
\begin{array}{c}
\begin{array}{ccc} \mathrm{A} & \mathrm{B} & \mathrm{D} \end{array} \\
\begin{bmatrix} 0.00 & 2.70 & 0.30 \\ 2.70 & 0.00 & 3.00 \\ 0.30 & 3.00 & 0.00 \end{bmatrix}
\end{array}
$$

এখানে A ও D হতে B-এর গড় দূরত্ব [2.85] বেশি। সুতরাং {B} হবে Splinter গ্রুপ এবং {A, B} মূল গ্রুপ বলে বিবেচিত হবে। এখন Splinter গ্রুপ হতে অন্যান্য এককের গড় দূরত্ব এবং মূল গ্রুপের গড় দূরত্ব হবে নিম্নরূপ :

একক	Splinter গ্রুপ হতে গড় দূরত্ব	মূল গ্রুপের গড় দূরত্ব	উভয় দূরত্বের পার্থক্য
A	2.70	0.30	−2.4
D	3.00	0.30	−2.7

এখানে উভয় দূরত্বের পার্থক্য ঋণাত্মক হওয়াতে আর কোনো একক Splinter গ্রুপে যুক্ত হবে না। সুতরাং {A, B, D} বিভক্ত হয়ে {B} ও {A, D} গুচ্ছদ্বয় হবে। অনুরূপভাবে প্রয়োজন অনুসারে অপর গুচ্ছকেও Splint করা যেতে পারে এবং n এককে n গুচ্ছে বিভক্ত করা যায়। তবে, এই পদ্ধতিতে বা Agglomerative পদ্ধতিতে গুচ্ছায়ন কোন পর্যায়ে শেষ হবে তা নির্ধারণ করা প্রয়োজন। এ সম্পর্কে গুচ্ছায়নের মূল্যায়ন করা বা গুচ্ছায়ন পরিসংখ্যানিকভাবে তাৎপর্য কিনা তা যাচাই করে দেখতে হয়। এই যাচাই বা গুচ্ছায়নের মূল্যায়ন ৭.৪ অনুচ্ছেদে করা হবে। তার আগে Agglomerative পদ্ধতি সম্পর্কে আরো কিছু তথ্য আলোচনা করা যাক।

Agglomerative পদ্ধতির অনেকগুলো সমস্যা আছে। তার মধ্যে প্রথম হলো, এই পদ্ধতিতে কোনো একটি একক একবার কোনো গুচ্ছভুক্ত হলে তা অন্য কোনো গুচ্ছে আর অন্তর্ভুক্ত হবে না। এককের প্রথম গুচ্ছভুক্তি ভুলবশত হলেও তা আর অন্য গুচ্ছে অন্তর্ভুক্ত হবে না। দ্বিতীয়ত, এই পদ্ধতিতে এককসমূহ শৃঙ্খলিত হয়। অর্থাৎ নতুন একক পুরাতন গুচ্ছের সাথে শৃঙ্খলিত হয়ে গুচ্ছভুক্ত হবে কিন্তু নতুন একক নিয়ে নতুন গুচ্ছ হবে না। অবশ্য, এ ধরনের সমস্যা গুচ্ছায়নের প্রথম দিকেই হয়। গুচ্ছায়ন শৃঙ্খলিত পদ্ধতিতে হয় বলে গুচ্ছসমূহ কিছুটা অসদৃশ হয়ে থাকে। তাছাড়া Single linkage এবং Comple lingage ছাড়া অন্য পদ্ধতিতে গুচ্ছায়ন করা হলে তা দূরত্ব ম্যাট্রিক্স-এর মানের পরিবর্তনের সাথে পরিবর্তিত হয়ে থাকে। আগেই আলোচনা করা হয়েছে যে দূরত্ব ম্যাট্রিক্স-এর মান চলকের মানের এককের পরিবর্তনের সাথে পরিবর্তিত হয়ে থাকে।

৭.৩.৩ Partitioning পদ্ধতি : এই পদ্ধতির মূল বৈশিষ্ট্য হলো যে একবার কোনো একক ভুলক্রমে কোনো গুচ্ছে অন্তর্ভুক্ত হলে তা আবার অন্য গুচ্ছভুক্ত হতে পারে। এই পদ্ধতিতে এককসমূহ কোনো পূর্ব নির্ধারিত নির্দেশক অনুসরণে

গুচ্ছভুক্ত হয় এবং গুচ্ছের সংখ্যা আগে থেকেই নির্ধারিত থাকে। অবশ্য গুচ্ছায়ন করার সময় গুচ্ছের সংখ্যার পরিমাণে পরিবর্তিত হতে পারে।

K-গড় গুচ্ছায়ন (K-means Clustering) : ধরা যাক n এককের একটি নমুনা আছে এবং প্রতি নমুনা একক হতে P-চলকের মান পরিমাপ করা হয়েছে। ধরা যাক $X_{ij}(i=1, 2, \cdots, n\ ;\ j=1, 2, \cdots, p)$ হলো i-তম একক হতে পরিমাপ করা j-তম চলকের মান। অনুমান করা যাক যে প্রতিটি চলকের জন্য Euclidean দূরত্ব পরিমাপ করা যায়। ধরা যাক এককসমূহ K গুচ্ছে অন্তর্ভুক্ত হবে এবং ধরা যাক k গুচ্ছে অন্তর্ভুক্ত হওয়ার কারণে এককসমূহ $P_{n,k}$ ভাগে বিভক্ত হবে। এখানে K-এর মান অনেকভাবে নির্ণয় করা যায়। যেমন, (1) প্রথম K এককে প্রাথমিক K গুচ্ছ গড় ভেক্টর বিবেচনা করা যায়, (2) K-এর পূর্ব নির্ধারিত কোনো মান বিবেচনা করা যায়, (৩) যে সকল একক K এককের মধ্যে দূরত্ব বেশি সে সকল একক নিয়ে K গুচ্ছ বিবেচনা করা যায়।

K-এর মান পূর্ব নির্ধারিত হলে i-তম একক কোন গুচ্ছভুক্ত হবে তা নির্ণয় করার জন্য নিম্নলিখিত সূত্র প্রয়োগ করা যায় :

$$K(i) = \frac{K\{\text{Sum}(i) - \text{Min}[\text{Sum}(i)]\}}{\text{Max}[\text{Sum}(i)] - \text{Min}[\text{Sum}(i)]} + 1$$

এখানে Sum(i) হলো i-তম এককের ক্ষেত্রে সকল চলকের মানের যোগফল। এই সূত্র প্রয়োগ করে যে সকল এককের জন্য K(i)-এর একই গোটা মান পাওয়া যাবে সেগুলো এক গুচ্ছভুক্ত হবে। এই গুচ্ছই হলো প্রাথমিক গুচ্ছ।

ধরা যাক l-তম গুচ্ছে $(l=1, 2, \cdots, k)$ এককের সংখ্যা হলো n_l এবং $\bar{X}_{lj}$ হলো l-তম গুচ্ছে অন্তর্ভুক্ত j-তম $(j=1, 2, \cdots, p)$ চলকের গড়। অর্থাৎ

$$\bar{X}_{lj} = \frac{1}{n_l} \sum_{i=1}^{n_l} X_{ij}\ ,\quad l=1, 2, \ldots, k$$

কাজেই i-তম একক এবং l-তম গুচ্ছের মধ্যে দূরত্ব হবে

$$d_{il} = \left[\sum_{j=1}^{p} \{ X_{ij} - \bar{X}_{lj} \}^2 \right]^{1/2}$$

তাহলে $P_{n,k}$ ভাগে বিভক্ত করার কারণে বিচ্যুতি হবে

$$E[P_{n,k}] = \sum_{i=1}^{n} [d_{i(i)}]^2$$

এখানে $l(i)$ হলো l-তম গুচ্ছ যার মধ্যে i-তম একক আছে, $d_{i(i)}$ হলো i-তম একক এবং j-তম একক যে l-তম গুচ্ছে অন্তর্ভুক্ত আছে সে গুচ্ছ গড়ের মধ্যে Euclidean দূরত্ব এবং $E[P_{n,k}]$ হলো n এককে k গুচ্ছায়নের মাধ্যমে বিভক্ত করার কারণে বিচ্যুতি।

এখন গুচ্ছায়ন করার ক্ষেত্রে এককসমূহকে এমনভাবে বিভক্ত করতে হবে যেন $E[P_{n,k}]$-এর মান কম হয় এবং যে কোনো এককে এক গুচ্ছ থেকে অন্য গুচ্ছে এমনভাবে স্থানান্তরিত করতে হবে যে স্থানান্তরের ফলে $E[P_{n,k}]$-এর ন্যূনতম মান পাওয়া যায়। প্রতিটি এককে এক গুচ্ছ হতে অন্য গুচ্ছে স্থানান্তরিত করতে হবে এবং স্থানান্তরের ফলে $E[P_{n,k}]$-এর মানের কি পরিবর্তন হয় তা লক্ষ্য করতে হবে। $E[P_{n,k}]$-এর মানের পরিবর্তন লক্ষ্য করার জন্য $R_{l(i),m}$ নামক তথ্যজমান নির্ণয় করতে হয়। এখানে

$$R_{l(i),m} = \frac{n_m d_{im}^2}{n_m + 1} - \frac{n_l d_{il}^2}{n_l - 1}, \quad l \neq m = 1, 2, \cdots, k$$ । এখানে

n_l = l-তম গুচ্ছে এককের সংখ্যা, n_m = m-তম গুচ্ছে এককের সংখ্যা। এই $R_{l(i),m} > 0$ হলে বুঝতে হবে যে i-তম এককে l-তম গুচ্ছ হতে m-তম গুচ্ছে স্থানান্তরের ফলে $E[P_{n,k}]$ এর মান বৃদ্ধি পায় এবং স্থানান্তর অকার্যকর কিন্তু $R_{l(i),m} < 0$ হলে স্থানান্তর যুক্তিসঙ্গত এবং নতুন করে গুচ্ছায়ন করতে হয়। এই নতুন গুচ্ছায়নের ফলে প্রাথমিক $E[P_{n,k}]$ এর মানে যে পরিবর্তন হয় তার পরিমাণ হলো

$$E[P'_{n,k}] = E[P_{n,k}] + R_{l(i),m}$$

এখানে $P'_{n,k}$ হলো n এককে নতুন করে k গুচ্ছে গুচ্ছায়ন করার ফলে এককের মোট ভাগের সংখ্যা এভাবে $R_{l(i),m}$ এর মান প্রতি এবং স্থানান্তরের পর করতে হবে। এই পদ্ধতি ততক্ষণ পর্যন্ত চলবে যতক্ষণ m এর সব মানের জন্য $R_{l(i),m} > 0$ হবে। বিষয়টি একটি উদাহরণের সাহায্যে ব্যাখ্যা করা যাক।

উদাহরণ ৭.১ ঃ উদাহরণ ৪.৪.১ হতে পাঁচটি একক দৈব পদ্ধতিতে চয়ন করে ঐ এককসমূহের তথ্য নিচে দেয়া হলো।

একক	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	Sum(i)	K(i)
A	2	1	0	0	3	0	4	0	5	15	1
B	8	1	1	0	10	1	1	1	19	42	3
C	5	2	0	0	3	2	4	0	12	28	1
D	3	1	0	0	3	0	4	0	11	22	1
E	4	2	0	0	3	1	4	0	11	25	1

ধরা যাক এই পাঁচটি এককে $k=2$টি গুচ্ছে বিভক্ত করতে হবে। তাহলে, প্রাথমিক গুচ্ছ দুটি হবে

গুচ্ছ-1 : (A, C, D, E) এবং গুচ্ছ-2 : (B) এবং গুচ্ছভুক্তি অনুসারে গুচ্ছ গড় হবে নিম্নরূপ :

গুচ্ছ	গুচ্ছে j-তম চলকের গড়								
	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
A, C, D, E	3.5	1.5	0	0	3	0.75	4	0	9.75
B	8	1	1	0	10	1	1	1	19

এখন l-তম গুচ্ছে অন্তর্ভুক্ত l-তম একক ও l-তম গুচ্ছের গুচ্ছ গড়ের মধ্যে Euclidean দূরত্বের বর্গ নির্ণয় করা যাক। এই দূরত্বসমূহ হলো

$$d_{11}^2 = (2-3.5)^2+(1-1.5)^2+(0-0)^2+(0-0)^2+(3-3)^2$$
$$+(0-0.75)^2+(4-4)^2+(0-0)^2+(5-9.75)^2$$
$$=25.625$$

$$d_{22}^2=(8-8)^2+(1-1)^2+(1-1)^2+(0-0)^2+(10-10)^2$$
$$+(1-1)^2+(1-1)^2+(1-1)^2+(19-19)^2$$
$$=0$$

$$d_{31}^2=(5-3.5)^2+(2-1.5)^2+(0-0)^2+(0-0)^2+(3-3)^2$$
$$+(2-0.75)^2+(4-4)^2+(0-0)^2+(12-9.75)^2$$
$$=9.125$$

$$d_{41}^2 = (3-3.5)^2+(1-1.15)^2+(0-0)^2+(0-0)^2+(3-3)^2$$
$$+(0-0.75)^2+(4-4)^2+(0-0)^2+(11-9.75)^2$$
$$=2.3975$$
$$d_{51}^2 = (4-3.5)^2+(2-1.5)^2+(0-0)^2+(0-0)^2+(3-3)^2$$
$$+(1-0.75)^2+(4-4)^2+(0-0)^2+(11-9.75)^2$$
$$=2.125$$

সুতরাং প্রাথমিক গুচ্ছায়নের বিচ্যুতি হলো

$$E[P_{5,2}] = \sum_{i=1}^{5} d_{i^2,\ (i)}$$
$$= d_{11}^2 + d_{22}^2 + d_{31}^2 + d_{41}^2 + d_{51}^2$$
$$= 39.2725$$

এখন A-কে দ্বিতীয় গুচ্ছে অন্তর্ভুক্ত করা হলে Euclidean দূরত্বের বর্গ পাওয়া যায়

$$d_{12}^2 = (2-8)^2+(1-1)^2+(0-1)^2+(0-0)^2+(3-10)^2$$
$$+(0-1)^2+(4-1)^2+(0-1)^2+(5-19)^2$$
$$=293$$

কাজেই $E[P_{5,2}]$-এর মানের পরিবর্তন হলো

$$R_{1(1),2} = \frac{n_2 d_{12}^2}{n_2+1} - \frac{n_1 d_{11}^2}{n_1-1}$$
$$= \frac{1\times 293}{1+1} - \frac{4\times 25.625}{4-1}$$
$$=112.33 > 0$$

লক্ষ্য করা যাচ্ছে যে A-কে দ্বিতীয় গুচ্ছে অন্তর্ভুক্ত করার ফলে E-এর মান বৃদ্ধি পাচ্ছে। সুতরাং A প্রথম গুচ্ছেই থাকবে। এখন C-কে দ্বিতীয় গুচ্ছে অন্তর্ভুক্ত করা হলে Euclidean দূরত্বের বর্গ পাওয়া যায়

$$d_{32}^2 = (5-8)^2+(2-1)^2+(0-1)^2+(0-0)^2+(3-10)^2$$
$$+(2-1)^2+(4-1)^2+(0-1)^2+(12-19)^2$$
$$=120$$

এবং $$R_{1(3),2} = \frac{n_2 d_{32}^2}{n_2+1} - \frac{n_1 d_{31}^2}{n_1-1} = 47.83 > 0$$

সুতরাং প্রথম গুচ্ছে থাকবে। অনুরূপভাবে পাওয়া যায়

$$d_{42}^2 = 149, \quad d_{52}^2 = 141, \quad R_{1(4),2} = 71.3 > 0$$

এবং $R_{1(5),2} = 67.67 > 0$। কাজেই D এবং F প্রথম গুচ্ছেই থাকবে। অর্থাৎ এক্ষেত্রে প্রাথমিক গুচ্ছই ফাইনাল গুচ্ছ হবে এবং প্রথম গুচ্ছ ও দ্বিতীয় গুচ্ছ গড়ের Euclidean দূরত্ব হবে 12.889।

Partitioning পদ্ধতি সম্পর্কে আরো তথ্য (More about Partitioning Method) : ধরা যাক $x_1, x_2, \ldots, x_n$ হলো n এককের একটি নমুনা, যেখানে x_i এর ($i = 1, 2, \ldots, n$) প্রতিটি p চলকের মানের ভেক্টর এবং প্রতিটি অনপেক্ষ। ধরা যাক এই n একক k সম্ভাব্য উপ-গণসমষ্টি হতে চয়ন করা নমুনা এবং l-তম ($l = 1, 2, \cdots, k$) উপ-গণসমষ্টির সম্ভাবনা ঘনত্ব ফাংশন হলো $f(x, \theta_l)$। এখানে k-এর মান অজানা।

ধরা যাক $x_j (j \neq i = 1, 2, \cdots, n)$ l-তম উপ-গণসমষ্টি হতে চয়ন করা হয়েছে কিনা তা চিহ্নিত করার জন্য একটি নির্দেশক (criterion) ভেক্টর হলো $\Gamma = (\gamma_1, \gamma_2, \cdots, \gamma_n)'$। এই $\gamma_j (i = 1, 2, \ldots, n)$ এর মান এমন যে $\gamma_j = l$ হলে বুঝতে হবে x_j l-তম উপ-গণসমষ্টি হতে চয়ন করা হয়েছে। ধরা যাক C_l হলো Γ-এর মাধ্যমে l-তম গুচ্ছে x_j এর সংখ্যা। এখন k গুচ্ছে অন্তর্ভুক্ত এককসমূহের জন্য সম্ভাব্যতা ফাংশন হলো

$$L(\Gamma ; \theta_1, \theta_2, \ldots, \theta_k) = \prod_{x \in C_1} f(x_1, \theta_1) \cdots \prod_{x \in C_k} f(x_k, \theta_k)$$

ধরা যাক $\Gamma, \theta_1, \theta_2, \cdots, \theta_k$ এর ML নিরূপক হলো $\hat{\Gamma}, \hat{\theta}_1, \cdots, \hat{\theta}_k$ যথাক্রমে। ধরা যাক $\hat{\Gamma}$-এর অধীনে $\hat{C}_l$ হলো l-তম গুচ্ছে x_j এর সংখ্যা। আরো ধরা যাক এককসমূহ স্থানান্তরিত হওয়ার ফলে $\hat{C}_l$ পরিবর্তিত হয়ে $\hat{C}_m (m \neq l = 1, 2, \cdots, k)$ হয়েছে। এই পরিবর্তনের ফলে সম্ভাব্যতা ফাংশন-এর আকার দাঁড়ায়

$$\frac{L(\hat{\Gamma}; \hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k) f(x, \hat{\theta}_m)}{f(x, \hat{\theta}_l)} \leqslant L(\hat{\Gamma}; \hat{\theta}_1, \cdots, \hat{\theta}_k)$$

সুতরাং $f(x, \hat{\theta}_m) \leqslant f(x, \hat{\theta}_l)$

কাজেই কোনো একক l-th গুচ্ছ হতে স্থানান্তরিত হয়ে m-th গুচ্ছভুক্ত হতে হলে $f(x, \hat{\theta}_m) < f(x, \hat{\theta}_l)$ হবে।

ধরা যাক $f(x, \theta_l)$ হলো $N_p(\mu_l, \Sigma_l)$, $l = 1, 2, \cdots, k$ এর সম্ভাবনা ঘনত্ব ফাংশন (p, d, f)। সেক্ষেত্রে

$$L(\Gamma ; \theta_1, \theta_2, \cdots, \theta_k) = \text{Const} - \tfrac{1}{2} \sum_{l=1}^{k} \sum_{x_i \in C_l} (x_i - \mu_l)' \Sigma_l^{-1} (x_i - \mu_l)$$

$$- \tfrac{1}{2} \sum_{l=1}^{k} n_l \log | \Sigma_l |$$

এখানে C_l-এ n_l একক আছে। এখন Γ দেয়া থাকলে μ এবং Σ-এর ML নিরূপকের জন্য L-এর মান বৃহত্তম হবে। অর্থাৎ Γ-এর জন্য

$$\hat{\mu}_l(\Gamma) = \bar{x}_l \quad \text{এবং} \quad \hat{\Sigma}_l(\Gamma) = S_l$$

এখানে $\bar{x}_l$ হলো C_l এর মধ্যে অন্তর্ভুক্ত n_l এককের গড় এবং S_l হলো ঐ এককসমূহের নমুনা সহ-ভেদাঙ্ক ম্যাট্রিক্স। এই $\hat{\mu}_l$ এবং S_l এর মান log L-এ বসিয়ে পাওয়া যায়

$$\log L \left[\Gamma ; \hat{\theta}(\Gamma)\right] = \text{Const} - \tfrac{1}{2} \Sigma n_l \log | S_l |$$

সুতরাং যে গুচ্ছায়ন $\prod_{l=1}^{k} | S_l |^{n_l}$ কে ন্যূনতম করবে তাই Γ-এর ML নিরূপক।

এভাবে C_l গুচ্ছে বিভক্ত করার জন্য জন্য l-th গুচ্ছে কমপক্ষে $(p+1)$ একক থাকা দরকার যেন $n_l \geq (p+1)$ হয় এবং $n \geq k(p+1)$ হয়। এ ধরনের গুচ্ছায়নের ক্ষেত্রে

$$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_l = \Sigma \text{ (অজানা)}$$

হলে উপরিউক্ত পদ্ধতিতে গুচ্ছায়ন $| W |$ -কে ন্যূনতম করবে। এখানে

$$W = \sum_{l=1}^{k} \sum_{C_l} (x_i - \bar{x}_l)(x_i - \bar{x}_l)'$$

এখানে x_j হলো C_i -এ অন্তর্ভুক্ত i-তম এককের মান। এই W গুচ্ছে অন্তর্ভুক্ত এককসমূহের প্রতিটি চলকের জন্য নির্ণয় করা যায় এবং সেক্ষেত্রে W-কে আন্তঃ গুচ্ছ বিচ্যুতি ম্যাট্রিক্স বলা হয়। ধরা যাক B হলো গুচ্ছ গড়ভিত্তিক বিচ্যুতি ম্যাট্রিক্স এবং T হলো সকল এককের জন্য বিচ্যুতি ম্যাট্রিক্স। তাহলে,

$$T = W + B$$

গুচ্ছায়নের ফলে বা গুচ্ছ থেকে একক স্থানান্তরের ফলে W বা B এর মানে পরিবর্তন হয় কিন্তু T-এর মানে কোনো পরিবর্তন হয় না বলে W বা B-এর ভিত্তিতে গুচ্ছায়ন নির্দেশক (criterion) নির্ণয় করা যায়। T-এর সংজ্ঞা অনুযায়ী W-এর মান কম হওয়ার অর্থ হলো B-এর মান বেশি হওয়া। B-এর মান বেশি হওয়ার অর্থ হলো গুচ্ছের এককসমূহের সদৃশ্যতা বৃদ্ধি পাওয়া এবং বিভিন্ন গুচ্ছের এককসমূহের মধ্যে বৈসদৃশ্যতা বৃদ্ধি পাওয়া। সুতরাং গুচ্ছায়ন করার নির্দেশক হিসেবে tr(W)-কে ন্যূনতম করা যেতে পারে। এছাড়া আরো দুটি নির্দেশক হলো (1) $|W|$-কে ন্যূনতম করা এবং (2) $\text{tr } BW^{-1}$ কে ন্যূনতম করা। এখানে $|W|$-কে ন্যূনতম করা বা $|T|/|W|$-কে বৃহত্তম করা একই কথা। Friedman and Rubin (1967) $|T|/|W|$-এর পরিবর্তে $\log[\max |T|/|W|]$ কে গুচ্ছায়নের নির্দেশক হিসেবে ব্যবহার করার প্রস্তাব করেছেন। জানা আছে যে,

$$\text{tr } BW^{-1} = \sum_{i}^{p} \lambda_i$$

এখানে $\lambda_1, \lambda_2, \ldots, \lambda_p$ হলো BW^{-1} এর আইগেন মান। সুতরাং $\Sigma\lambda_i$ ন্যূনতম করেও গুচ্ছায়ন করা যায়। এদিকে Scott and Symon (1971) $|W|$-কে ন্যূনতম করার পরিবর্তে $\prod_{l=1}^{k} |W_l|^{n_l}$-কে ন্যূনতম করে গুচ্ছায়ন করার প্রস্তাব করেছেন।

৭.৩.৪ গুচ্ছায়নের অন্যান্য পদ্ধতি (Other Methods of Clustering) : গুচ্ছায়ন সম্পর্কে Agglomerative ও Divisive পদ্ধতি আলোচনা করা হয়েছে। বিভিন্ন পদ্ধতি প্রয়োগের ক্ষেত্রে গুচ্ছায়ন সম্পর্কে সঠিক সিদ্ধান্ত নেয়ার সুবিধার্থে চিত্রের সাহায্য নেয়া যেতে পারে। চিত্রের সাহায্যে গুচ্ছায়ন করা হলে তা উপাত্ত সম্পর্কে সঠিক ধারণা ব্যক্ত করতে, উপাত্তের সঠিক বিশ্লেষণের পদক্ষেপ নিতে এবং উপাত্ত সম্পর্কে সঠিক সিদ্ধান্ত নিতে সাহায্য করে। Anderson (1954, 1957, 1960) p-মাত্রার চিত্রের মাধ্যমে গুচ্ছায়ন করার প্রস্তাব করেছেন। তিনি

যে চিত্রের কথা উল্লেখ করেছেন সেগুলো হলো (1) glyphs এবং (2) metro-glyphs। প্রতিটি উপাত্ত ভেক্টরকে একটি glyph দ্বারা প্রকাশ করা যায়। আর এই glyph হলো একটি বৃত্ত যার রশ্মিসহ একটি নির্দিষ্ট ব্যাসার্ধ আছে। এক একটি চলকের জন্য এক একটি রশ্মি ব্যবহার করা হয়ে থাকে। রশ্মির দৈর্ঘ্য চলকের মানের উপর নির্ভরশীল। Glyphs এবং metroglyphs ছাড়া আরো দুটি চিত্র মাধ্যম হলো Fourier series এবং Chernoff (1973) faces। এ সম্পর্কে বিস্তারিত জানার জন্য Anderson (1960), Andrews (1972), Flenberg (1979), Gnanadeskian (1977), Huff and Black (1978) আলোচনা করা যেতে পারে।

৭.৪ গুচ্ছায়ন সম্পর্কে যাচাই (Test Regarding Clustering)

গুচ্ছ তৈরি করার পর গুচ্ছগুলোর মধ্যে পার্থক্য কি রকম তা জানা দরকার। তাছাড়া, গুচ্ছের সঠিক সংখ্যা কত হওয়া উচিত অথবা গুচ্ছায়ন কতটা সঠিক হয়েছে সে সম্পর্কে সিদ্ধান্ত নেয়া উচিত। গুচ্ছায়নের সঠিকতা পর্যালোচনার একটি উপায় হলো l-তম গুচ্ছে j-তম চলকের ভেদাঙ্ক পর্যালোচনা করা। কোনো গুচ্ছে j-তম চলকের ভেদ যতো কম হবে গুচ্ছায়ন ততো ভাল হবে বলে সিদ্ধান্ত নেয়া যায়। তবে এ ব্যাপারে পরিসংখ্যানিক যাচাই পদ্ধতি সহজ নয়। কারণ, গুচ্ছায়ন (1) গুচ্ছ গড়ের সর্বোত্তম ভেদাঙ্ক (B, ম্যাট্রিক্স) (2) min tr (W) এবং (3) max ($|T|/|W|$)-এর উপর নির্ভরশীল। ফলে বিভিন্ন নির্দেশক হতে বিভিন্ন ফলাফল পাওয়া যায়। তবুও গুচ্ছায়ন সম্পর্কে একটি সিদ্ধান্ত নেয়ার জন্য গুচ্ছে অন্তর্ভুক্ত প্রতিটি চলকের জন্য ভেদাঙ্ক বিশ্লেষণ F যাচাই করা যায়। নিম্নে উদাহরণ ৭.১-এর ক্ষেত্রে গুচ্ছায়নের পর প্রতিটি চলকের আন্তঃগুচ্ছ ভেদাঙ্ক পর্যালোচনার জন্য F-যাচাই তথ্যজমান দেয়া হলো :

চলক	গুচ্ছ MS	d. f.	বিচ্যুতি MS	d. f.	F	P-value
y	16.20	1	1.6667	3	9.72	0.053
x_1	0.20	1	0.3333	3	0.60	0.495
x_2	0.80	1	0.0000	—	—	—
x_3	0.00	1	0.0000	—	—	—
x_4	0.20	1	0.0000	—	—	—
x_5	0.05	1	0.9167	3	0.05	0.830

x_5	7.20	I	0.0000	—	—	—
x_7	0.80	I	0.0000	—	—	—
x_8	68.45	I	10.25	3	6.68	0.081

লক্ষ্য করা যাচ্ছে যে, সকল চলকের ক্ষেত্রেই আন্তঃগুচ্ছ ভেদাঙ্ক কম এবং গুচ্ছ গড়ের ভেদাঙ্কই বেশি (যদিও উক্ত উদাহরণের ক্ষেত্রে তাৎপর্যপূর্ণ নয়)। এখানে আরো একটি লক্ষণীয় বিষয় হলো গুচ্ছায়নের ক্ষেত্রে চলক y, x_4 এবং x_8 অধিক প্রভাবশালী এবং প্রধানত এই তিনটি চলকের কারণেই চলক-B একটি ভিন্ন গুচ্ছে অন্তর্ভুক্ত হয়েছে। তাত্ত্বিকভাবে যাচাই সম্পর্কে আরো বিস্তারিত জানার জন্য Arnold (1979), Johnson (1972), Lee (1979), McClain and Rao (1975) এবং Wolfe (1970) পর্যালোচনা করা যেতে পারে।

গুচ্ছায়নের সঠিক সংখ্যা নির্ধারণ করার জন্য

$$\max \frac{|T|}{|W|}$$

এর উপর নির্ভর করা যেতে পারে। যে গুচ্ছায়নের ফলে $|T|/|W|$ সর্বোচ্চ হবে তাই গুচ্ছায়নের সঠিক সংখ্যা হিসেবে বিবেচিত হতে পারে।

গুচ্ছায়ন করার পর গুচ্ছসমূহের দূরত্ব নির্ণয় করা যায়। এই দূরত্ব হলো গুচ্ছ গড়সমূহের মধ্যে Euclidean দূরত্ব। এই দূরত্ব যতো বেশি হবে গুচ্ছায়ন ততো ভাল হবে বিবেচনা করা যায়। উদাহরণ ৭.১-এর ক্ষেত্রে এরূপ দূরত্ব হলো 12.889। এই দূরত্ব নির্ণয়ের সংখ্যা হলো

$$d_{lm} = \left[\sum_{j=1}^{p} (\bar{X}_{lj} - \bar{X}_{mj})^2 \right]^{1/2}$$

অষ্টম অধ্যায়

নির্ণায়ক বিশ্লেষণ
(Discriminant Analysis)

৮.১ সূচনা

বহুচলক বিশ্লেষণে একই নমুনা বিন্দু হতে একাধিক চলকের মান পরিমাপ করে বিভিন্ন পরিসংখ্যানিক বিশ্লেষণ করা হয়। ধরা যাক, সন্তান প্রসবে সক্ষম মায়েদের একটি নমুনা হতে প্রতি মায়ের বয়স, শিক্ষা, পেশা, মোট জীবিত প্রসব করা সন্তানের সংখ্যা, মোট মৃত সন্তানের সংখ্যা ইত্যাদি বিষয়ের উপর উপাত্ত সংগৃহীত হয়েছে। এখন মৃত সন্তানের সংখ্যার ভিত্তিতে মায়েদেরকে কতগুলো পরস্পর বৈসদৃশ (mutually exclusive) গুচ্ছে বিভক্ত করা যেতে পারে। এরূপ গুচ্ছায়ন করার একটি পদ্ধতি সপ্তম অধ্যায়ে 'গুচ্ছায়ন বিশ্লেষণ' এর অধীনে করা হয়েছে। ঐ গুচ্ছায়নের ক্ষেত্রে গুচ্ছ সম্পর্কে কোনো প্রাক (apriori) তথ্য জানা থাকে না। বস্তুসমূহের মধ্যে দূরত্ব বা সদৃশতা-এর ভিত্তিতে ঐগুলোকে গুচ্ছায়ন করা হয়। কিন্তু উপরিউক্ত উদাহরণের ক্ষেত্রে মৃত সন্তানের সংখ্যা সম্পর্কে প্রাক তথ্য জানা থাকে এবং এই জানা তথ্যের ভিত্তিতে মায়েদেরকে গুচ্ছায়ন করে গুচ্ছসমূহের পার্থক্য পর্যালোচনা করা যেতে পারে। এই পার্থক্য পর্যালোচনা করার জন্য নমুনা বিন্দুগুলোর অনপেক্ষ চলকসমূহের একটি রৈখিক সমাবেশ নির্ণয় করা হয় এবং এই রৈখিক সমাবেশই প্রাক তথ্যের ভিত্তিতে একটি নমুনা বিন্দু সঠিক গুচ্ছে অন্তর্ভুক্ত কিনা সে সম্পর্কে ধারণা পেতে সাহায্য করে। প্রাক তথ্যের ভিত্তিতে নির্ণয় করা রৈখিক সমাবেশ দ্বারা একটি নমুনা বিন্দু কোন গুচ্ছে অন্তর্ভুক্ত হবে সে সম্পর্কে ধারণা লাভ করাই হলো নির্ণায়ক বিশ্লেষণের (Discriminant analysis) কাজ। সুতরাং নির্ণায়ক বিশ্লেষণ হলো কতগুলো অনপেক্ষ চলকের ভিত্তিতে নমুনা বস্তুসমূহকে (objects) পরস্পর বৈসদৃশ এবং সর্বসংবলিত (mutually exclusive and exhaustive) শ্রেণিতে শ্রেণিবিন্যাস করার একটি পরিসংখ্যানিক পদ্ধতি। Sir Ronald Fisher সর্বপ্রথম এই বিশ্লেষণ উপস্থাপন করেন।

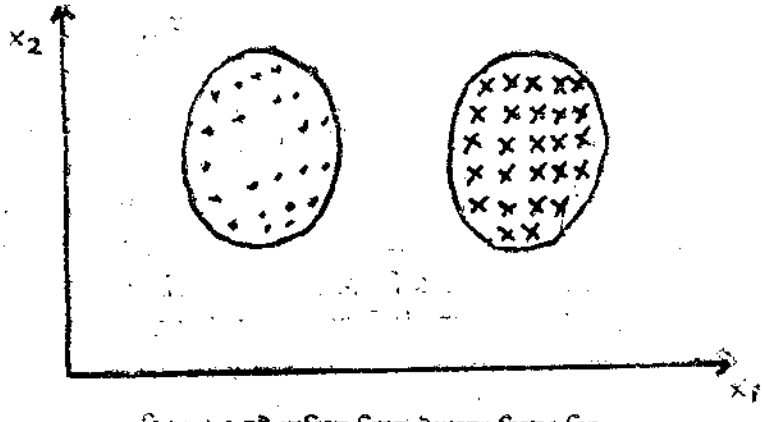
৮.২ নির্ণায়ক বিশ্লেষণের কর্মধারা (Scope of Discriminant Analysis)

নির্ণায়ক বিশ্লেষণের মাধ্যমে বস্তুসমূহকে দুই গুচ্ছ বা দুই-এর অধিক গুচ্ছে বিভক্ত করা হয়। এখানে রৈখিক সমাবেশ কোনো নমুনা বিন্দুকে যে কোনো একটি গুচ্ছে অন্তর্ভুক্ত হওয়া সম্পর্কে সিদ্ধান্ত নিতে সাহায্য করে। এরূপ সিদ্ধান্ত নেয়ার

মৌলিক নীতি (basic principle) হলো, যে কোনো বস্তুর ক্ষেত্রে ভুলবশত শ্রেণি-ভুক্ত হওয়ার (misclassification) বিচ্যুতি (error) যেন কম হয়। এই মৌলিক নীতি বহাল থাকলে বস্তুসমূহের অন্তঃগুচ্ছ (within group) ভেদাঙ্ক-এর তুলনায় আন্তঃগুচ্ছ (between group) ভেদাঙ্ক বেশি হয়। এখন প্রশ্ন হলো অন্তঃগুচ্ছ ভেদাঙ্ক কম হওয়ার পরিসংখ্যানিক তাৎপর্য কি? এককসমূহকে গুচ্ছে বা শ্রেণিতে বিভক্ত করা হলে এবং বিভক্তিকরণে বিচ্যুতি কম হলে অন্তঃগুচ্ছ ভেদাঙ্ক কম হবে এবং শ্রেণিভুক্তকরণ সঠিক হলে অন্তঃগুচ্ছ ভেদাঙ্ক শূন্য-এর কাছাকাছি হবে। বিভক্তিকরণে যতো বেশি বিচ্যুতি হবে যে কোনো চলকের মোট ভেদাঙ্ক ততো বেশি হবে এবং অন্তঃগুচ্ছ বর্গসমষ্টি ও মোট বর্গ সমষ্টি-এর অনুপাত ততো বড় হবে। এই অনুপাত নির্ণয় করা হয় U-তথ্যজমান দ্বারা বা Wilk's Lamba (Λ) দ্বারা।

গুচ্ছসমূহের গুচ্ছ গড়গুলো সমান হলে Λ-এর মান 1 হবে। অর্থাৎ অন্তঃগুচ্ছ ভেদাঙ্ক মোট ভেদাঙ্কের সমান এবং শ্রেণিভুক্তকরণে বিচ্যুতি বেশি। Λ-এর মান শূন্য হলে বুঝতে হবে আন্তঃগুচ্ছ ভেদাঙ্ক ও মোট ভেদাঙ্ক সমান। এক্ষেত্রে শ্রেণি-ভুক্তকরণে বিচ্যুতি কম।

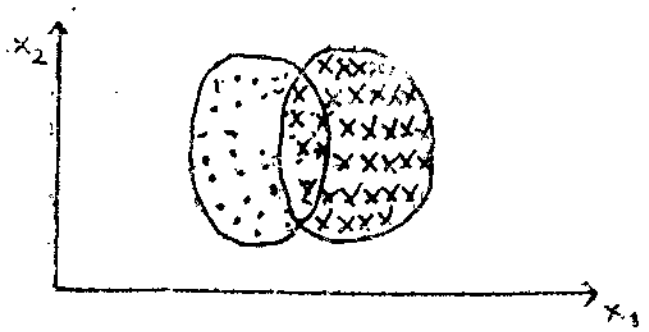
ধরা যাক কোনো একটি নমুনার এককসমূহকে দুটি শ্রেণিতে বিভক্ত করা হলো। এখন দুই শ্রেণির উপাত্ত সত্যিকারভাবে দুই গণসমষ্টিভুক্ত হলে এবং প্রতিটি নমুনা হতে দুটি চলকের মান পরিমাপ করা হলে চলকদ্বয়ের ভিত্তিতে অঙ্কিত বিক্ষেপ চিত্রের আকার হবে চিত্র ৮.১-এর ন্যায়। এখানে x_1 ও x_2 হলো চলকদ্বয়।



চিত্র ৮.১ : দুই শ্রেণিতে বিভক্ত উপাত্তের বিক্ষেপ চিত্র।

আবার, ধরা যাক দুই শ্রেণিতে বিভক্ত হলেও উপাত্তসমূহের মধ্যে কিছু মিল আছে। সেক্ষেত্রে বিক্ষেপ চিত্রের আকার হবে চিত্র ৮.২-এর ন্যায়। এক্ষেত্রে নির্ণায়ক বিশ্লেষণের কাজ হলো অনপেক্ষ চলকসমূহের এমন একটি রৈখিক সমাবেশ নির্ণয়

করা যা পূর্ব নির্ধারিত গুচ্ছসমূহের মধ্যে এমনভাবে পার্থক্য নির্দেশ করবে (Discriminate) যেন পার্থক্য নির্দেশ করার ক্ষেত্রে বিচ্যুতি হার কম হয়।



চিত্র ৮.২ : দুই শ্রেণিতে বিভক্ত কিন্তু কিছু একের সঙ্গে মিল আছে এমন উপাত্তের বিক্ষেপ চিত্র।

উপরিউক্ত রৈখিক সমাবেশকে লেখা যায় $D = b'x$, এখানে D হলো $(1 \times n)$ আকারের নির্ণায়ক সাফল্যাঙ্ক [Discriminant Score] এর ভেক্টর, b হলো $(p \times 1)$ আকারের নির্ণায়ক [Discriminant] ভর এর একটি ভেক্টর এবং X হলো $n \times p$ আকারের উপাত্ত ম্যাট্রিক্স।

নির্ণায়ক বিশ্লেষণ করার জন্য নমুনা এককসমূহকে গুচ্ছে বিভক্ত করার সুবিধার্থে একটি যুগ্ম (binary) চলক বা নির্দেশক (Indicator) চলক ব্যবহার করা হয়। এখানে যুগ্ম চলকের কাজ হলো একটি গুচ্ছের এককসমূহকে চিহ্নিত করার জন্য যুগ্ম চলকের মান 0 ধরা হয় এবং অপর গুচ্ছের এককসমূহকে চিহ্নিত করা হয় যুগ্ম চলকের মান 1 দ্বারা। এই যুগ্ম চলকের প্রসঙ্গে X-এর মানের ভিত্তিতে একটি রৈখিক সমাবেশ $D = b'x$ নির্ণয় করা এবং D-এর মানের ভিত্তিতে নমুনা এককসমূহকে শ্রেণিভুক্ত করা হয়। এখানে সাফল্যাঙ্ক D হলো b নিরূপিত হওয়ার পর যে কোনো নমুনা এককের চলকসমূহের মান নিয়ে গঠিত ভেক্টর ও b'-এর গুণফল। এদিক থেকে বিবেচনা করা হলে নির্ণায়ক বিশ্লেষণ এবং নির্ভরণ বিশ্লেষণের মধ্যে একটি মিল আছে। কেননা যুগ্ম চলকের ক্ষেত্রে এই চলককে নির্ভরশীল চলক (y) ধরে এবং x-এর চলকসমূহকে অনপেক্ষ চলক বিবেচনা করা হলে $D = b'x$ এর ন্যায় রৈখিক সম্ভাবনা মডেল [Linear probability model] $Y = b'x$ পাওয়া যায়। কিন্তু তা সত্ত্বেও নির্ভরণ বিশ্লেষণ ও নির্ণায়ক বিশ্লেষণ এক নয়। তাদের মধ্যে ফাংশনের সদৃশ্যতা থাকলেও বৈসদৃশতাই বেশি।

প্রথমত নির্ভরণ বিশ্লেষণের ক্ষেত্রে অনুমান করা হয় যে y-চলক পরিমিত বিন্যাস অনুসরণ করে এবং x-এর চলকসমূহ কোনো পরিসংখ্যানিক বিন্যাস অনুসরণ

করে না। কিন্তু নির্ণায়ক বিশ্লেষণের ক্ষেত্রে যুগ্ম চলক (Binary variable) কোনো পরিসংখ্যানিক বিন্যাস অনুসরণ করে না। অপরপক্ষে X হলো $N_p(\mu, \Sigma)$ হতে চয়ন করা একটি উপাত্ত ম্যাট্রিক্স। দ্বিতীয়ত নির্ভরণ বিশ্লেষণের কাজ হলো অনপেক্ষ চলকের মানের ভিত্তিতে y-চলক সম্পর্কে পূর্বাভাস করা। অপরপক্ষে, নির্ণায়ক বিশ্লেষণের কাজ হলো অনপেক্ষ চলকের রৈখিক সমাবেশ নির্ণয় পূর্বক একক-সমূহকে নির্ভুলভাবে বা ন্যূনতম বিচ্যুতিসহ শ্রেণিভুক্ত করা। এছাড়া নির্ভরণ বিশ্লেষণের ক্ষেত্রে নির্ভরশীল চলকের জন্য একটি মডেল বিবেচনা করে ঐ মডেল হতে যে কোনো অনপেক্ষ চলকের প্রতি একক পরিবর্তনের জন্য নির্ভরশীল চলকের কি পরিমাণ পরিবর্তন হচ্ছে তা নিরূপণ করা। কিন্তু নির্ণায়ক বিশ্লেষণের ক্ষেত্রে একক-সমূহকে নির্ভুলভাবে শ্রেণিভুক্ত করার একটি উপায় বের করা হয়।

৮.৩ নির্ণায়ক বিশ্লেষণের ক্ষেত্রে অনুমান (Assumptions Underlying Discriminant Analysis)

আগেই উল্লেখ করা হয়েছে যে, নির্ণায়ক বিশ্লেষণের কাজ হলো অনপেক্ষ চলকের রৈখিক সমাবেশ নির্ণয় করে তার ভিত্তিতে নমুনা এককসমূহকে নির্ভুলভাবে শ্রেণিভুক্ত করা। এই শ্রেণিভুক্তকরণের কাজ প্রকৃষ্টভাবে (optimally) করার জন্য উপাত্ত সম্পর্কে কিছু অনুমান করতে হয়। এগুলো হলো :

(১) অনপেক্ষ চলকসমূহ বহুচলক পরিমিত বিন্যাস অনুসরণ করে।

(২) বিভিন্ন শ্রেণির উপাত্ত হতে প্রাপ্ত $(p \times p)$ ভেদাঙ্ক সহভেদাঙ্ক ম্যাট্রিক্সগুলো সমমাত্রিক (homogeneous) হবে।

উপরিউক্ত প্রথম অনুমান যাচাই করার জন্য p চলকের প্রতিটি পরিমিত বিন্যাস অনুসরণ করে কিনা তা লক্ষ্য করা যেতে পারে। কারণ p-চলক বহুচলক পরিমিত বিন্যাস অনুসরণ করলে তাদের প্রতিটি এক-চলক পরিমিত বিন্যাস অনুসরণ করে। কাজেই কোনো চলক পরিমিত বিন্যাস অনুসরণ না করলে উপাত্ত ম্যাট্রিক্স বহুচলক পরিমিত বিন্যাস অনুসরণ করবে না। সেক্ষেত্রে ঐ চলককে বিশ্লেষণ হতে বাদ দেয়া যেতে পারে। অবশ্য প্রতিটি চলক ভিন্ন ভিন্নভাবে পরিমিত বিন্যাস অনুসরণ করলেও তাদের যুগ্ম বিন্যাস বহুচলক পরিমিত বিন্যাস অনুসরণ করবে এমন কোনো কথা নেই। বহুচলক পরিমিত বিন্যাসের যাচাই সম্পর্কে আরো বিস্তারিত জানার জন্য Andrews (1972) পর্যালোচনা করা যেতে পারে।

বিভিন্ন শ্রেণির সহভেদাঙ্ক ম্যাট্রিক্সগুলোর সমমাত্রিকতা যাচাই করারও বিভিন্ন পদ্ধতি আছে। ঐগুলোর মধ্যে একটি পদ্ধতি হলো Box's (1947) M-test (3.2.26)। এই যাচাই এর মাধ্যমে $H_0 : \Sigma_1 = \Sigma_2 = \cdots, = \Sigma_k$ বাতিল হওয়ার অনেক কারণ থাকতে পারে। তার মধ্যে একটি হলো উপাত্ত ম্যাট্রিক্স পরিমিত বিন্যাস

অনুসরণ না করা এবং অপরটি হলো নমুনা আকার বড় হওয়া। নমুনা আকার বড় হলে সহভেদাঙ্ক ম্যাট্রিক্সগুলো খুব বেশি বিসদৃশ না হলেও নাস্তিকল্পনা বাতিল হয়ে যেতে পারে। কারণ বড় নমুনার ক্ষেত্রে তাৎপর্য সম্ভাবনা (Significance probability মোটামুটি সদৃশ সহভেদাঙ্ক ম্যাট্রিক্স-এর ক্ষেত্রেও ছোট হতে পারে। যে যাই হোক অসমমাত্রিক সহ-ভেদাঙ্ক ম্যাট্রিক্স-এর ক্ষেত্রে নির্ণয় বিশ্লেষণ করা হয় দ্বিঘাত নির্ণায়ক ফাংশনের (Quadratic discriminant function) সাহায্যে (Wahl Kronmal, 1977)।

৮.২ নির্ণায়ক বিশ্লেষণের যৌক্তিকতা (Justification of Discriminant Analysis)

ধরা যাক k গণসমষ্টি হতে k নমুনা চয়ন করা হয়েছে এবং নমুনা এককসমূহ হতে p চলকের মান পরিমাপ করা হয়েছে। আরো ধরা যাক যে প্রতিটি গণসমষ্টি ভিন্ন ভিন্ন বিন্যাস অনুসরণ করে। মনে করি l-তম ($l=1, 2, \cdots, k$) গণসমষ্টির সম্ভাবনা ঘনত্ব ফাংশন হলো $f_l(x)$। অর্থাৎ l-তম গণসমষ্টি হতে একটি নমুনা একক চয়ন করা হলে ঐ নমুনা এককের জন্য প্রাপ্ত চলকসমূহের ঘনত্ব ফাংশন হবে $f_l(x)$। এখানে x হলো p চলকের মানের ভেক্টর। এখন নির্ণায়ক বিশ্লেষণের কাজ হলো যে কোনো একটি নমুনা এককের x-এর মান জানা থাকলে ঐ নমুনা একক কোনো গণসমষ্টিভুক্ত তা চিহ্নিত করা। চিহ্নিত করার এ কাজ এমনভাবে করতে হয় যেন চিহ্নিতকরণ পদ্ধতির বিচ্যুতি কম হয়।

একটি উদাহরণের সাহায্যে বিষয়টি ব্যাখ্যা করা যাক। ধরা যাক একজন ডাক্তারের নিকট অনেক রোগী এসেছে এবং তাদের প্রত্যেকের রোগের ধরন বিভিন্ন। ডাক্তারের কাজ হলো রোগের লক্ষণ পর্যালোচনা করে রোগীর রোগ নির্ণয় করা এবং সে মোতাবেক চিকিৎসার ব্যবস্থা করা। ডাক্তারকে রোগ নির্ণয় করার সময় নির্ভুল সিদ্ধান্ত নিতে হয়। রোগ নির্ণয়ে বিচ্যুতি যতো কম হবে রোগীর চিকিৎসা ততো সহজ হবে। আবার ধরা যাক কোনো সরকারী দপ্তরে লোক নিয়োগ করা হবে এবং সে উদ্দেশ্যে প্রার্থীর বিভিন্ন ধরনের পরীক্ষা নেয়া হয়েছে। এক্ষেত্রে সকল প্রার্থীর ক্ষেত্রেই একইরূপ তথ্য পরিমাপ করা হয়েছে। এখন ঐ তথ্যসমূহের ভিত্তিতে সঠিকভাবে প্রার্থী বাছাই করতে হবে। এখানে প্রার্থী চিহ্নিত করার জন্য সরকারী কর্মকর্তাকে এমন কিছু নির্দেশক নির্ণয় করতে হবে যেন ন্যূনতম বিচ্যুতির মাধ্যমে প্রার্থী নিয়োগ করা যায়। নির্ণায়ক বিশ্লেষণের কাজ হলো এমন একটি নির্দেশক নির্ণয় করা যা প্রার্থীদেরকে দুই বা ততোধিক গুচ্ছে বিভক্ত করতে পারে।

ধরা যাক প্রার্থীদের প্রত্যেকের ক্ষেত্রে চলক $x_1, x_2, \cdots, x_p$-এর পরিমাপ নথিভুক্ত করা হয়েছে এবং প্রার্থীদেরকে দুটি গুচ্ছে বিভক্ত করা হবে। ধরা যাক গুচ্ছ দুটি হলো :

(1) উপযুক্ত প্রার্থীদের গুচ্ছ এবং

(2) অনুপযুক্ত প্রার্থীদের গুচ্ছ। এখন প্রথমোক্ত গুচ্ছকে 1 দ্বারা এবং শেষোক্ত গুচ্ছকে 0 দ্বারা চিহ্নিত করা হলে নির্ণায়ক বিশ্লেষণের কাজ হলো চলক $x_1, x_2, \cdots, x_p$-এর এমন একটি রৈখিক সমাবেশ নির্ণয় করা যা নির্ভুলভাবে যে কোনো প্রার্থীকে সঠিক গুচ্ছে অন্তর্ভুক্ত করতে পারে। এরূপ একটি রৈখিক সমাবেশ হলো

$$D = \beta_0 + \beta_1 x_1 + \beta_2 x_1 + \cdots + \beta_p x_p$$

ধরা যাক একটি গুচ্ছে n_1 এবং অপর গুচ্ছে n_2 একক আছে এবং নমুনায় $n_1 + n_2 = n$ একক আছে। ধরা যাক

$$\bar{X}_1 = [\bar{x}_{11}\ \bar{x}_{12} \cdots \bar{x}_{1p}]' \text{ এবং } \bar{X}_2 = [\bar{x}_{21}\ \bar{x}_{22} \cdots \bar{x}_{2p}]'$$

এখন নমুনা দুটি একই গণসমষ্টিভুক্ত হলে $\mu_1 = \mu_2$ হবে, এখানে μ_1 এবং μ_2 হলো $\bar{X}_1$ ও $\bar{X}_2$ এর প্রাসঙ্গিক গণসমষ্টি গড়। সুতরাং $H_0 : \mu_1 = \mu_2$ বাতিল হলে আলোচিত উপাত্ত ব্যবহার করে নির্ণায়ক বিশ্লেষণ করা যেতে পারে। এই নাস্তি-কল্পনা যাচাই পদ্ধতি ৩.২.৩ অনুচ্ছেদে আলোচনা করা হয়েছে। নমুনা k গণসমষ্টি হতেও চয়ন করা যেতে পারে। সেক্ষেত্রে $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ যাচাই করা যেতে পারে। এই যাচাই পদ্ধতি ৯.৩ অনুচ্ছেদে আলোচনা করা হয়েছে।

আবার, ধরা যাক μ_{lj} $(l = 1, 2, \cdots, k\ ;\ j = 1, 2, \cdots, p)$ হলো j-তম চলকের l-তম নমুনার গণসমষ্টি গড়। এখানে

$$H_0 : \mu_{1j} = \mu_{2j} = \cdots = \mu_{kj}$$

এই নাস্তিকল্পনা যাচাই একমুখী শ্রেণিবিন্যাসের জন্য ব্যবহৃত F-যাচাই-এর মাধ্যমে করা যায়। নাস্তিকল্পনা বাতিল হলে নির্ণায়ক বিশ্লেষণের যৌক্তিকতা উপলব্ধি করা যায়। কারণ k গণসমষ্টি ভিন্ন ভিন্ন হলেই নাস্তিকল্পনা বাতিল হবে।

উপরিউক্ত আলোচনা হতে বুঝা যাচ্ছে যে, কিছু বর্ণনামূলক পরিসংখ্যান (Descriptive statistics) এবং এক চলক (Univariate) যাচাই নির্ণায়ক বিশ্লে-ষণ করার যৌক্তিকতা ব্যাখ্যা করতে পারে। অবশ্য নির্ণায়ক বিশ্লেষণের ক্ষেত্রে একচলক বিশ্লেষণ যথেষ্ট নয়। সব চলক একসঙ্গে নিয়েই এই বিশ্লেষণ করা হয়। তবে, একচলক বিশ্লেষণের মাধ্যমে j-তম চলকের ক্ষেত্রে

$$H_0 : \mu_{1j} = \mu_{2j} = \ldots = \mu_{kj}$$

তাহা হলে j-তম চলক নির্ণায়ক বিশ্লেষণে অন্তর্ভুক্ত করা তেমন ফলদায়ক হবে না।

নির্ণায়ক বিশ্লেষণ করার পর নির্ণায়ক ফাংশনের (Discriminant function) মূল্যায়ন করা হয়ে থাকে। এই মূল্যায়ন সম্ভাবনা তত্ত্বের উপর নির্ভরশীল। এ সম্পর্কে আলোচনা ৮.৪.৪ অনুচ্ছেদে করা হবে। এই ফাংশন হলো

$$D = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

একে নির্ভরণ মডেলের ন্যায় দেখা যাচ্ছে। এখানে D হলো নির্ণায়ক সাফল্যাঙ্ক (Discriminant score) নির্ণায়ক ফাংশন নির্ণয় করার পর প্রতি একক হতে প্রাপ্ত চলকের মানের ভিত্তিতে D-এর মান নির্ণয় করা যায়। এখন চলকসমূহ নির্ণায়ক ফাংশন নির্ণয়ে কতটা অবদান রাখে তা নির্ণয় করার জন্য D-এর মান ও x_j এর মানের সংশ্লেষাঙ্ক নির্ণয় করা হয়। এই সংশ্লেষাঙ্কের পরিমাণ এবং চিহ্ন চলক-সমূহের সংশ্লেষাঙ্ক দ্বারা প্রভাবিত হয়। বিষয়টি উদাহরণ ৮.১ এর উপাত্তের ক্ষেত্রে ব্যাখ্যা করা হবে। যা হোক, এটি বুঝা যাচ্ছে যে নির্ণায়ক বিশ্লেষণের আগে অন্তঃগুচ্ছ চলকসমূহের সংশ্লেষাঙ্ক নির্ণয় করে সকল নমুনা এককের জন্য একত্রিত অন্তঃগুচ্ছ (Pooled within group) সংশ্লেষণ ম্যাট্রিক্স নির্ণয় করা প্রয়োজন, কারণ কোনো জোড়া চলক বেশি সংশ্লেষিত হলে D-এর সাথে তাদের যে কোনো-টির সংশ্লেষাঙ্ক কোনো অর্থবহ তথ্য সরবরাহ করবে না। সুতরাং বুঝা যাচ্ছে যে নির্ণায়ক বিশ্লেষণ হতে যুক্তিসঙ্গত তাৎপর্য ব্যাখ্যা করার জন্য চলকসমূহের সংশ্লে-ষণ পর্যালোচনা করা প্রয়োজন।

কোনো জোড়া চলক বেশি সংশ্লেষিত হওয়ার অর্থই হলো তাদের মধ্যে একটি রৈখিক সম্পর্ক আছে। এরূপ রৈখিক সম্পর্ক যে কোনো একটি অনপেক্ষ চলকের সাথে অন্য সব চলকের হতে পারে। অর্থাৎ একটি অনপেক্ষ চলক অন্য অনপেক্ষ চলকসমূহের রৈখিক সমাবেশ। ধরা যাক j-তম চলকের সাথে অন্যান্য চলকের বহুল সংশ্লেষাঙ্ক (Multiple correlation co-efficient) হলো R_j। তাহলে j-তম চলকের সাথে অন্যান্য রৈখিক সম্পর্কের মাত্রা পরিমাপ করা যেতে পারে $(1 - R_j^2)$ দ্বারা। এই $(1 - R_j^2)$ পরিমাপকে বলা হয় টলারেন্স (Tolerance)। যে সব চলকের জন্য টলারেন্স-এর মান ছোট ঐগুলোকে নির্ণায়ক বিশ্লেষণে অন্তর্ভুক্ত করা যুক্তিসঙ্গত নয়।

৮.৪ নির্ণয়ন পদ্ধতি (Method of Discrimination)

ধরা যাক $p_1, p_2, \ldots, p_k$ গণসমষ্টি আছে এবং তাদের মাঝিক বিন্যাস ও আছে। এখন ঐ গণসমষ্টির যে কোনো একটি হতে একটি এককের বিভিন্ন চলকের পরিমাপ

পাওয়া গেলে তা কোন গণসমষ্টিভুক্ত হবে তা নির্ধারণ করাই নির্ণায়ক বিশ্লেষণের মুখ্য উদ্দেশ্য। এই নির্ধারণ প্রক্রিয়ার পদ্ধতিগুলো হলো :

(1) সর্বোত্তম সম্ভাব্য নির্ণায়ক পদ্ধতি (Maximum likelihood discriminant rule)

(2) Bayes নির্ণায়ক পদ্ধতি (Bayes discriminant rule)

(3) Fisher-এর রৈখিক নির্ণায়ক ফাংশন (Fisher's linear discriminant function)।

৮.৪.১ সর্বোত্তম সম্ভাব্য নির্ণায়ক পদ্ধতি (Maximum likelihood discriminant rule) : ধরা যাক $p_1, p_2, \cdots, p_k$-এর যে কোনো একটি হতে একটি এককের উপাত্ত x পাওয়া গেছে। তাহলে সর্বোত্তম সম্ভাব্য নির্ণায়ক পদ্ধতি হলো ঐ নমুনা এককটিকে $p_i\,(i=1, 2, \cdots, k)$ গণসমষ্টিতে অন্তর্ভুক্ত করা যেতে পারে যদি

$$L_i(x) = \max_i L_i(x)$$

এখানে $L_i(x)$ হলো x-এর মানের জন্য যে সম্ভাব্যতা ফাংশন (Likelihood function) বৃহত্তম হবে। কিন্তু x-এর মানের জন্য অনেকগুলো সম্ভাব্যতা ফাংশন বৃহত্তম হলে তাদের যে কোনো একটিকে x-এর গণসমষ্টি হিসেবে চিহ্নিত করা যেতে পারে।

ধরা যাক $k=2$ এবং $P_1 \sim N(\mu_1, \sigma_1^2)$; $P_2 \sim N(\mu_2, \sigma_2^2)$। তাহলে

$$L_i(x) = (2\pi\sigma_1^2)^{-1/2} \exp\left\{\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right\} \; ; \; i=1, 2$$

এখন $L_1(x) > L_2(x)$ হবে যদি

$$\frac{\sigma_2}{\sigma_1} \exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \left(\frac{x-\mu_2}{\sigma_2}\right)^2\right]\right\} > 1$$

হয়। উভয় পাশে লগ নিয়ে এবং সরল করে পাওয়া যায়

$$x^2\left[\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right] - 2x\left[\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right] + \left[\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right] < 2\log\frac{\sigma_2}{\sigma_1}$$

এখন $\mu_2 > \mu_1$ এবং $\sigma_1 > \sigma_2$ হলে x-এর খুব ছোট মান ও খুব বড় মানের জন্য উপরিউক্ত অসমতা বহাল থাকে। কিন্তু $\sigma_1 = \sigma_2$ হলে

$$|x-\mu_2| > |x-\mu_1|$$

এর ক্ষেত্রে $L_1(x) > L_2(x)$ হবে। সুতরাং $\mu_2 > \mu_1$ হলে x P_2-এর অন্তর্ভুক্ত হবে যদি

$$x > \tfrac{1}{2}(\mu_1+\mu_2)$$

হয়। অন্যথায় x, P_1 এর অন্তর্ভুক্ত হবে। কিন্তু $\mu_1 > \mu_2$ হলে এবং

$$x > \tfrac{1}{2}(\mu_1+\mu_2)$$

এর ক্ষেত্রে x, P_1 এর অন্তর্ভুক্ত হবে।

ধরা যাক $P_1, P_2, \ldots, P_k$ এর বিন্যাস জানা আছে কিন্তু এদের মানসমূহ অজানা। সেক্ষেত্রে পরামানসমূহ নিরূপণ করতে হয়। ধরা যাক $X' = (X_1', X_2', \ldots, X_k')$, যেখানে X_i হলো i-তম গণসমষ্টি হতে চয়ন করা $(n_i \times p)$ আকারের উপাত্ত ম্যাট্রিক্স এবং X হলো $(n \times p)$ $[\Sigma n_i = n]$ আকারের উপাত্ত ম্যাট্রিক্স।

ধরা যাক $P_i \sim N_p(\mu_i, \Sigma)$। সেক্ষেত্রে $\mu_1, \mu_2, \ldots, \mu_k$ এর নিখুঁতি নিরূপক হলো যথাক্রমে $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$ এবং Σ-এর নিখুঁতি নিরূপক হলো $S = \Sigma n_i S_i/(n-k)$, এখানে $\bar{X}_i$ এবং S_i হলো যথাক্রমে i-তম নমুনার নমুনা গড় এবং নমুনা ভেদাঙ্ক। এখন $k=2$ হলে এবং x যে কোনো নমুনা বিন্দুর চলকের পরিমাপ হলে সর্বোত্তম সম্ভাব্য নির্ণায়ক পদ্ধতি অনুসারে x গণসমষ্টি P_1 -এ অন্তর্ভুক্ত হবে এবং কেবল যদি

$$b'\{x - \tfrac{1}{2}(\bar{X}_1 + \bar{X}_2)\} > 0 \text{ হয়,}$$

এখানে $b = S^{-1}(\bar{X}_1 - \bar{X}_2)$। বিষয়টি একটি উদাহরণের সাহায্যে ব্যাখ্যা করা যাক।

উদাহরণ ৮.৯ : নিচে Mahagub (1996)-এর কাজ হতে 35 দম্পতির জীবিত জন্মগ্রহণ করা সন্তানের সংখ্যা ও আরো কিছু তথ্য উপস্থাপন করা হলো। যেখানে x_1 = জীবিত জন্মগ্রহণ করা সন্তানের সংখ্যা, x_2 = দম্পতির বিবাহিত জীবনকাল, x_3 = আকাঙ্খিত সন্তানের সংখ্যা, x_4 = মায়ের শিক্ষা, x_5 = পিতার পেশা, $[x_5 = 1$, শিক্ষকতা ; $x_5 = 2$, অফিস কর্মী ; $x_5 = 3$, ব্যবসা ; $x_5 = 4$, কৃষিকাজ$]$, x_6 = মায়ের পেশা $[x_6 = 1$, গৃহিনী ; $x_6 = 2$, অন্যান্য$]$।

ক্রমিক নম্বর	x_1	x_2	x_3	x_4	x_5	x_6
1	3	14	3	6	1	1
2	2	8	2	14	3	1
3	4	18	4	7	1	1

4	3	10	3	10	1	1
5	2	5	2	14	3	1
6	2	6	2	14	3	1
7	3	7	3	10	3	1
8	5	15	5	8	1	1
9	4	16	4	7	1	1
10	3	10	3	10	1	1
11	4	15	4	7	1	1
12	3	8	3	10	1	1
13	3	12	3	10	3	2
14	2	8	2	10	3	2
15	2	4	2	14	3	2
16	2	7	3	14	3	2
17	4	10	4	8	4	2
18	4	14	4	6	4	2
19	3	12	3	6	4	2
20	2	8	2	5	4	2
21	3	10	3	0	4	2
22	2	12	2	5	4	2
23	3	9	3	6	4	2
24	4	16	4	8	4	2
25	2	5	2	6	1	2
26	4	18	4	7	1	2
27	3	10	3	6	1	2

28	4	12	4	7	1	2
29	2	8	2	8	1	2
30	1	5	2	10	1	2
31	1	3	2	5	1	2
32	3	8	3	8	1	2
33	0	2	2	8	1	2
34	0	1	2	6	1	2
35	1	3	2	6	1	2

উপরিউক্ত উপাত্তের ক্ষেত্রে x_6 এর মানের ভিত্তিতে নমুনাকে দুটি ভাগে বিভক্ত করা হলে প্রথম দুটি চলকের (x_1, x_2) ভিত্তিতে একটি নির্ণায়ক বিশ্লেষণ করা যাক। ধরা যাক $x_1 = 5$ এবং $x_2 = 18$ অন্য একটি নমুনা বিন্দু হতে পর্যবেক্ষণ করা হয়েছে। এই নমুনা বিন্দুটি কোন গণসমষ্টিভুক্ত হবে তা চিহ্নিত করাই নির্ণায়ক বিশ্লেষণের কাজ।

এখানে x_1 ও x_2 এর ভিত্তিতে পাওয়া যায়

$$\bar{X}_1 = \begin{bmatrix} 3.17 \\ 11.00 \end{bmatrix}, \quad \bar{X} = \begin{bmatrix} 2.39 \\ 8.57 \end{bmatrix}, \quad S_1 = \begin{bmatrix} 0.8788 & 3.4546 \\ 3.4546 & 19.2727 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} 1.5217 & 4.8597 \\ 4.8597 & 19.9842 \end{bmatrix}, \quad n_1 = 12, \; n_2 = 23$$

Box's M-যাচাই দ্বারা বলা যায় যে S_1 ও S_2 সমমাত্রিক। কারণ $MC^{-1} = 1.92$। এই MC^{-1} এর মান 3 স্বাধীনতার মাত্রাবিশিষ্ট χ^2 এর সারণিকৃত মান অপেক্ষা ছোট।

এখন সর্বোত্তম সম্ভাব্য নির্ণায়ক নিয়ম অনুসারে পাওয়া যায়

$$b = S^{-1}(\bar{X}_1 - \bar{X}_2)$$

$$= \begin{bmatrix} 3.0223 & -0.6721 \\ -0.6721 & 0.2001 \end{bmatrix} \begin{bmatrix} 0.78 \\ 2.43 \end{bmatrix}$$

$$= \begin{bmatrix} 0.7242 \\ -0.0380 \end{bmatrix}$$

$$b'\left[x-\tfrac{1}{2}(\bar{X}_1+\bar{X}_2)\right]$$

$$=[0.7242-0.0380]\left[\begin{pmatrix}5\\18\end{pmatrix}-\begin{pmatrix}2.780\\9.785\end{pmatrix}\right]$$

$$=1.29>0$$

সুতরাং $x_1=5$ ও $x_2=18$ তথ্যজমান প্রথম গণসমষ্টির, অর্থাৎ $x_6=1$ উপাত্তবিশিষ্ট গণসমষ্টির একটি নমুনা।

এতক্ষণ দুই গণসমষ্টির যে কোনো একটি হতে নমুনা উপাত্ত পাওয়া গেলে ঐ নমুনা উপাত্ত কোন্ গণসমষ্টিভুক্ত হবে তা চিহ্নিত করার সর্বোত্তম সম্ভাব্য নির্ণায়ক পদ্ধতি আলোচনা করা হয়েছে। ধরা যাক তিনটি গণসমষ্টি হতে নমুনা চয়ন করা হয়েছে বিশ্লেষণ করার জন্য। সেক্ষেত্রে নির্ণায়ক ফাংশনকে লেখা যায়

$$D_{ij}=(\bar{X}_i-\bar{X}_j)'S^{-1}X-\tfrac{1}{2}\bar{X}_i'S^{-1}\bar{X}_i+\tfrac{1}{2}\bar{X}_j'S^{-1}\bar{X}_j$$

এখানে $i\neq j=1, 2, 3$; $\bar{X}_i$ এবং $\bar{X}_j$ হলো যথাক্রমে i-তম ও j-তম গণসমষ্টি হতে চয়ন করা নমুনা গড় ভেক্টর ; $S=\Sigma n_iS_i/(n-k)$; S_i হলো n_i নমুনা আকারবিশিষ্ট i-তম গণসমষ্টি হতে চয়ন করা নমুনার নমুনা ভেদাঙ্ক ; X হলো যে কোনো গণসমষ্টি হতে পাওয়া একটি নমুনা বিন্দুর উপাত্ত।

এই X-কে কোনো বিশেষ গণসমষ্টিভুক্ত করার নিয়ম হলো

$D_{12}>0$ এবং $D_{13}>0$ হলে X P_1 ভুক্ত হবে

$D_{12}<0$ এবং $D_{23}>0$ হলে X P_2 ভুক্ত হবে

$D_{13}<0$ এবং $D_{23}<0$ হলে X P_3 ভুক্ত হবে

বিষয়টি উদাহরণ ৮.১-এর উপাত্তের ক্ষেত্রে ব্যাখ্যা করা যাক। এখানে উপাত্তকে চলক x_5 এর মানের ভিত্তিতে তিনটি ভাগে বিভক্ত করা যাক। ধরা যাক $x_1=4$, $x_2=9$ এবং $x_6=10$ একটি নমুনা বিন্দু হতে পাওয়া গেছে। প্রশ্ন হলো এই নমুনা বিন্দুটি কোন গণসমষ্টিভুক্ত হবে ?

আলোচিত উদাহরণের ক্ষেত্রে $n_1=19$, $n_2=8$ এবং $n_3=8$।

$\bar{X}_1=[2.632 \quad 9.526 \quad 7.474]'$, $\bar{X}_2=[2.250 \quad 7.125 \quad 12.500]'$,

$\bar{X}_3=[3.125 \quad 11.375 \quad 5.500]'$

$$S_1=\begin{bmatrix}2.1345 & 7.3158 & 0.2953\\7.3158 & 29.7076 & 0.4035\\0.2953 & 0.4035 & 2.4854\end{bmatrix}$$

$$S_2 = \begin{bmatrix} 0.2143 & 0.6786 & -0.7143 \\ 0.6786 & 5.8393 & -3.2143 \\ -0.7143 & -3.2143 & 4.2857 \end{bmatrix}$$

$$S_3 = \begin{bmatrix} 0.6964 & 1.2321 & 0.9286 \\ 1.2321 & 7.1250 & 2.5000 \\ 0.9286 & 2.5000 & 6.2857 \end{bmatrix}$$

$$S^{-1} = \begin{bmatrix} 2.698996 & -0.622819 & -0.140791 \\ -0.622819 & 0.191616 & 0.031779 \\ -0.140791 & 0.031779 & 0.250159 \end{bmatrix}$$

এখন

$$h_{12} = (\bar{X}_1 - \bar{X}_2)' S^{-1} \{X - \tfrac{1}{2}(\bar{X}_1 + \bar{X}_2)\}$$

$$= 0.405$$

$$h_{13} = (\bar{X}_1 - \bar{X}_3)' S^{-1} \{X - \tfrac{1}{2}(\bar{X}_1 + \bar{X}_3)\}$$

$$= 1.237$$

$$h_{23} = (\bar{X}_2 - \bar{X}_3)' S^{-1} \{X - \tfrac{1}{2}(\bar{X}_1 + \bar{X}_2)\}'$$

$$= 0.807$$

এখানে $X = [4 \;\; 9 \;\; 10]'$। অনুমান করা হয়েছে যে S_1, S_2, S_3 সমমাত্রিক। লক্ষ্য করা যাচ্ছে যে $h_{12} > 0$ এবং $h_{13} > 0$। সুতরাং X $x_5 = 1$ মানবিশিষ্ট গণসমষ্টিভুক্ত।

৮.৪.২ সম্ভাব্য অনুপাত নির্ণায়ক পদ্ধতি (The Likelihood Ratio Discriminant Rule) : ধরা যাক X হলো i-তম গণসমষ্টির একটি নমুনা বিন্দু হতে পাওয়া চলকসমূহের মান নির্দেশকারী একটি ভেক্টর। মনে করি X_1 এবং X_j হলো যথাক্রমে P_1 এবং P_j গণসমষ্টি হতে চয়ন করা n_1 এবং n_j আকারের নমুনা হতে প্রাপ্ত উপাত্ত ম্যাট্রিক্স ($i \neq j = 1, 2, \ldots, k$)। এখানে নির্ণায়ক বিশ্লেষণের কাজ হলো X কোন গণসমষ্টিভুক্ত হবে তা নির্ণয় করা।

ধরা যাক নাস্তিকল্পনা হলো H_1 : X, P_j হতে চয়ন করা এবং X হলো X_j-এর একটি সারি। সর্বোত্তম সম্ভাব্য (ML) নির্ণায়ক পদ্ধতি অনুসারে X, P_j-এর অন্তর্ভুক্ত হবে, যদি X-এর মানের জন্য i-তম গণসমষ্টির বৃহত্তম সম্ভাব্য ফাংশন পাওয়া

যায়। Anderson (1958) এই পদ্ধতির বিকল্প হিসেবে H_1-এর অধীনে সম্ভাব্য কাংশন নির্ণয় করে তাদের অনুপাত নির্ণয় করার প্রস্তাব করেছেন। ধরা যাক $\hat{\Sigma}_i$ এবং $\hat{\Sigma}_j$ হলো যথাক্রমে H_i ও H_j $(i \neq j = 1, 2, \ldots, k)$ এর অধীনে সহ-ভেদাঙ্ক ম্যাট্রিক্স-এর সর্বোত্তম সম্ভাব্য নিরূপক [MLE]। এখন সম্ভাব্য অনুপাত তথ্যজমান হলো λ_i, যেখানে

$$\lambda_i = \frac{|\hat{\Sigma}_j|}{|\hat{\Sigma}_i|}$$

এই λ_i-এর মান 1-এর বেশি হলে H_i সত্য বলে বিবেচিত হবে এবং X, P_1-এর অন্তর্ভুক্ত হবে।

বিষয়টি দুটি গণসমষ্টির জন্য ব্যাখ্যা করা যাক। ধরা যাক $P_i \sim N_p(\mu_i, \Sigma_i)$ এবং $P_2 \sim N_p(\mu_2, \Sigma_2)$ অনুমান করা যাক যে $\Sigma_1 = \Sigma_2 = \Sigma$। তাহলে H_1-এর অধীনে μ_1, μ_2 এবং Σ-এর সর্বোত্তম সম্ভাব্য নিরূপক হলো যথাক্রমে

$$(n_1 \bar{X}_1 + X)/(n_1 + 1),\ \bar{X}_2 \text{ এবং}$$

$$\hat{\Sigma}_1 = \frac{1}{n_1 + n_2 + 1}\left\{ W + \frac{n_1}{1 + n_1}(X - \bar{X}_1)(X - \bar{X}_2)' \right\}$$

এখানে $W = n_1 S_1 + n_2 S_2$; $\bar{X}_i$ এবং S_i হলো i-তম গণসমষ্টি হতে চয়ন করা নমুনার যথাক্রমে নমুনা গড় ভেক্টর এবং নমুনা সহ-ভেদাঙ্ক ম্যাট্রিক্স; n_i হলো নমুনা আকার $(i = 1, 2)$। অনুরূপভাবে H_2-এর অধীনে উপরিউক্ত নিরূপকসমূহ হলো

$$\bar{X}_1,\ (n_2 \bar{X}_2 + X)/(n_2 + 1),$$

$$\hat{\Sigma}_2 = \frac{1}{n_1 + n_2 + 1}\left\{ W + \frac{n_2}{1 + n_2}(X - \bar{X}_2)(X - \bar{X}_2)' \right\}$$

সুতরাং, সম্ভাব্য অনুপাত তথ্যজমান λ হলো

$$\lambda = \left[\frac{|\hat{\Sigma}_2|}{|\hat{\Sigma}_1|} \right]^{\frac{1}{2}(n_1 + n_2 + 1}$$

$$= \left[\frac{1 + \frac{n_2}{1 + n_2}(X - \bar{X}_2)' W^{-1} (X - \bar{X}_2)}{1 + \frac{n_1}{1 + n_1}(X - \bar{X}_1)' W^{-1} (X - \bar{X}_1)} \right]^{\frac{n_1 + n_2 + 1}{2}}$$

Anderson উল্লেখ করেছেন যে $|\hat{\Sigma}_2| / |\hat{\Sigma}_1| > 1$ হলে X গণসমষ্টি P_1-এর অন্তর্ভুক্ত হবে। অর্থাৎ

$$\frac{n_2}{n_2+1}(X-\bar{X}_2)' W^{-1} (X-\bar{X}_2) > \frac{n_1}{n_1+1}(X-\bar{X}_1)' W^{-1} (X-\bar{X}_1)$$

হলেই H_1 গ্রহণযোগ্য হবে এবং X P_1-এর অন্তর্ভুক্ত হবে।

উপরিউক্ত সম্ভাব্য অনুপাত তথ্যজমান পদ্ধতি সর্বোত্তম সম্ভাব্য পদ্ধতির সমতুল্য যদি n_1 এবং n_2 সমান হয়। আলোচিত পদ্ধতিদ্বয় অভিসারীভাবেও সমতুল্য যদি n_1 ও n_2 বড় হয়। কিন্তু $n_1 \neq n_2$ হলে X P_1-এর অন্তর্ভুক্ত হওয়ার সম্ভাবনা থাকে যদি n_i বড় হয় $(i=1, 2)$।

ধরা যাক উদাহরণ ৮.১-এর ক্ষেত্রে $x_1 = 5$ এবং $x_2 = 18$। অর্থাৎ $X = [5 \quad 18]'$। এখানে

$$W^{-1} = \begin{bmatrix} 0.08649 & -0.01918 \\ -0.01918 & 0.00570 \end{bmatrix}$$

সুতরাং $$\frac{n_1}{n_1+1}(X-\bar{X}_1)' W^{-1} (X-\bar{X}_1) = 0.98$$

এবং $$\frac{n_2}{n_2+1}(X-\bar{X}_2) W^{-1} (X-\bar{X}_2) = 1.96$$

হওয়ায় সিদ্ধান্ত গ্রহণ করা যায় যে X P_1-এর অন্তর্ভুক্ত।

৮.৪.৩ Fisher-এর রৈখিক নির্ণায়ক ফাংশন (Fisher's Linear Discriminant Function) : ধরা যাক $P_1, P_2, \cdots, P_k$ গণসমষ্টি আছে এবং i-তম গণসমষ্টি হতে n_i নমুনা চয়ন করা হয়েছে, $i = 1, 2, k$। আরো ধরা যাক $X_i = [X_{i1}, X_{i2}, \ldots, X_{ip}]'$ হলো i-তম গণসমষ্টি হতে প্রাপ্ত $(n_i \times p)$ উপাত্ত ম্যাট্রিক্স এবং X হলো $(n \times p)$ $[\Sigma n_i = n]$ আকারের উপাত্ত ম্যাট্রিক্স। তাহলে সকল n উপাত্তের জন্য

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{X})(x_{ij} - \bar{X})'$$

হলো মোট বর্গসমষ্টি ম্যাট্রিক্স। i-তম গণসমষ্টি প্রাপ্ত নমুনা হতে এরূপ বর্গসমষ্টি হলো

$$W_i = \sum_{j=1}^{n_i} (x_{ij} - \bar{X}_i)(x_{ij} - \bar{X})'$$

তাহলে অন্তঃগুচ্ছ (within-groups) বর্গসমষ্টি হলো $W = W_1 + W_2 + \ldots + W_k$
এখন আন্তঃগুচ্ছ (between-groups) বর্গসমষ্টি ম্যাট্রিক্স হলো

$$B = T - W$$

ধরা যাক $D = b'X$ হলো রৈখিক নির্ণায়ক ফাংশন, এখানে D-কে বলা হয় রৈখিক কম্পোজিট (Linear composite)। D এমন একটি ফাংশন যা গণসমষ্টি-সমূহের মধ্যে পার্থক্য নির্দেশকারী নির্দেশক (criterion)। এই D-এর প্রাসঙ্গিক আন্তঃগুচ্ছ (Between-groups) বর্গসমষ্টি এবং অন্তঃগুচ্ছ (Within groups) বর্গ-সমষ্টি ম্যাট্রিক্সদ্বয় হলো যথাক্রমে $b'Bb$ এবং $b'Wb$। Fisher প্রস্তাব করেছেন $D = b'X$ এমনভাবে পেতে হবে যেন আন্তঃগুচ্ছ বর্গসমষ্টি ও অন্তঃগুচ্ছ বর্গসমষ্টি-এর অনুপাত বৃহত্তম হয়। ধরা যাক এই অনুপাত হলো

$$\lambda = \frac{b'Bb}{b'Wb}$$

এখন b এমনভাবে পেতে হবে যেন λ বৃহত্তম হয়। অর্থাৎ b এমনভাবে পেতে হবে যেন $(d\lambda/db) = 0$ হয়। এই সমীকরণ হতে সরলীকরণের মাধ্যমে পাওয়া যায়

$$(B - \lambda W)\, b = 0$$

অথবা

$$(W^{-1}B - \lambda I)\, b = 0$$

শেষোক্ত সমীকরণ হতে বলা যায় যে λ এর বৃহত্তম মান হবে $W^{-1}B$-এর বৃহত্তম আইগেন মান এবং b হলো ঐ আইগেন মানের প্রাসঙ্গিক আইগেন ভেক্টর।

আমরা পাই $X = [X_1\ X_2\ \cdots\ X_p]'$। এখন $X_1, X_2, \ldots, X_p$ রৈখিক-ভাবে অনপেক্ষ হলে এবং $n - k \geq p$ হলে rank $(W) = p$। আবার, rank (B) হলো p এবং $k - 1$ এর ন্যূনতম মান। ধরা যাক $W^{-1}B$-এর ক্রমসংখ্যা (rank) হলো r, যেখানে $r = \min(p, k-1)$।

এখানে $W^{-1}B$ প্রতিসম (Symmetric) ম্যাট্রিক্স নয়। $W^{-1}B$-এর উপর কোনো রোটেশন করা না হলে তার আইগেন ভেক্টরসমূহ অপরস্পরিত হলেও সম-কৌণিক (Orthogonal) হয় না।

ধরা যাক $W^{-1}B$-এর r সংখ্যক ভিন্ন ভিন্ন আইগেন মান আছে $(\lambda_1 > \lambda_2 > \cdots > \lambda_r)$। সুতরাং নির্ণায়ক বিশ্লেষণের মাধ্যমে r রৈখিক কম্পোজিট (Linear composite) পাওয়া যাবে। ধরা যাক λ_j হলো $W^{-1}B$ এর j-তম আইগেন মান। তাহলে এই আইগেন মান দ্বারা j-তম রৈখিক কম্পোজিটের গুরুত্ব পরিমাপ করা যেতে পারে। এরূপ একটি পরিমাপ হলো $\lambda_j / \Sigma\lambda_j$ এবং এটি দ্বারা গণসমষ্টিসমূহের

মধ্যে পার্থক্য নির্দেশ করতে কোন নির্ণায়ক ফাংশন অধিক ভেদের ব্যাখ্যা করতে পারে তা পরিমাপ করা যায়।

আলোচিত বিশ্লেষণের ক্ষেত্রে λ_j হলো $W^{-1}B$-এর আইগেন মান। এই আইগেন মানের প্রাসঙ্গিক আইগেন ভেক্টর হলো b_j। এখানে b_j-কে বলা হয় নির্ণায়ক ভর (Discriminant weights)। এই ভর অনপেক্ষ চলকের ভেক্টর X দ্বারা প্রভাবিত হয়ে থাকে। এ সম্পর্কে আলোচনা ৮.৩ অনুচ্ছেদে করা হয়েছে। সে কারণেই নির্ণায়ক ফাংশন ও অনপেক্ষ চলকসমূহের সম্পর্ক পর্যালোচনা করার জন্য এবং নির্ণায়ক বিশ্লেষণের ফলাফল ব্যাখ্যা করার জন্য যে কোনো রৈখিক কম্পোজিটের মান (D_i) ও অনপেক্ষ চলকের মানের সরল সংশ্লেষাঙ্ক নির্ণয় করতে হয়। এই সরল সংশ্লেষাঙ্কই গণসমষ্টিসমূহের মধ্যে পার্থক্য নির্দেশ করার ক্ষেত্রে j-তম চলকের গুরুত্ব ব্যাখ্যা করে।

নির্ণায়ক ফাংশন নির্ণয় করার পর যে কোনো এককের p চলকের মানের ভিত্তিতে ঐ একক কোন গণসমষ্টিভুক্ত হবে তার সিদ্ধান্ত নিতে হয়। ধরা যাক x হলো ঐ এককের চলকসমূহের মানের ভেক্টর। আবার, $\bar{X}_i$ হলো i-তম গণসমষ্টি হতে চয়ন করা নমুনা হতে প্রাপ্ত গড় ভেক্টর। কাজেই নির্ণায়ক সাফল্যাঙ্ক-এর গড় হবে $\bar{D}_i = b'\bar{X}_i$। এখন $b'x$ যদি $\bar{D}_i$ এর কাছাকাছি হয়, তাহলে x i-তম গণসমষ্টিভুক্ত হবে। অন্যভাবে বলা যায় যে x i-তম গণসমষ্টিভুক্ত হবে যদি

$$| b'x - b'\bar{X}_i | < | b'x - b'\bar{X}_l |, \quad i \neq l$$

Fisher-এর নির্ণায়ক ফাংশন $k = 2$ এর ক্ষেত্রে খুবই গুরুত্বপূর্ণ। সেক্ষেত্রে rank $(B) = 1$ এবং $B = (n_1 n_2/n)\, dd'$, যেখানে $d = \bar{X}_1 - \bar{X}_2$। সুতরাং $W^{-1}B$-এর শূন্য নয় এমন একটি আইগেন মান পাওয়া যায়। এই আইগেন মান হলো

$$\text{tr}\, W^{-1}B = \frac{n_1 n_2}{n} d'W^{-1}B, \quad n = n_1 + n_2$$

এবং এর প্রাসঙ্গিক আইগেন ভেক্টর হলো

$$b = W^{-1}d$$

সুতরাং x p_1 ভুক্ত হবে যদি

$$d'W^{-1}\{x - \tfrac{1}{2}(\bar{X}_1 + \bar{X}_2)\} > 0$$

হয়। অন্যথায় x p_2-এর অন্তর্ভুক্ত হবে।

লক্ষ্য করা যাচ্ছে যে এই নির্ণায়ক পদ্ধতি সর্বোত্তম সম্ভাব্য নির্ণায়ক পদ্ধতির সমতুল যদি $P_1 \sim N_p(\mu, \Sigma)$ এবং $P_2 \sim N_p(\mu, \Sigma)$ হয়। কিন্তু সর্বোত্তম সম্ভাব্য নির্ণায়ক পদ্ধতির ক্ষেত্রে $P_i \sim N_p(\mu_i, \Sigma)$ হতেই হয়। অপরপক্ষে Fisher-এর নির্ণায়ক পদ্ধতির জন্য P_i সঠিকভাবে $N_p(\mu_i, \Sigma)$ না হলেও চলে।

৮.৪.৪ Bayes নির্ণায়ক পদ্ধতি (Bayes Discriminant Rule) : ধরা যাক k গণসমষ্টি $p_1, p_2, \ldots, p_k$ আছে। ধরা যাক P_i এর প্রাক সম্ভাবনা (Prior probability) হলো $p(P_i)$। এই $p(P_i)$ নির্ণয় করার নানান পদ্ধতি আছে। ধরা যাক নমুনা উপাত্তসমূহ গণসমষ্টির প্রতিনিধিত্বকারী। তাহলে i-তম গণসমষ্টি হতে চয়ন করা এককের সমানুপাত $[n_i/n,\ n=\Sigma n_i]$ প্রাক সম্ভাবনার একটি নিরূপক হতে পারে। যেমন, ৮.১ উদাহরণের ক্ষেত্রে উপাত্তসমূহকে x_6-এর মানের ভিত্তিতে গুচ্ছায়ন করা হলে $n_1=12$ এবং $n_2=23$ হয়। এখানে প্রথম গুচ্ছ উপাত্তের জন্য $p(p_1)=0.34$ এবং দ্বিতীয় গুচ্ছ উপাত্তের জন্য $p(p_2)=0.66$।

অনেক সময় সব গুচ্ছই সমান সম্ভাব্য। সেক্ষেত্রে প্রতি গুচ্ছের জন্য প্রাক সম্ভাবনা সমান বিবেচনা করা যেতে পারে। ধরা যাক একটি নমুনার এককসমূহকে পুরুষ ও মহিলা হিসেবে চিহ্নিত করে দুটি গুচ্ছে ভাগ করা হলো। তাহলে, পুরুষ নমুনা এককের জন্য প্রাক সম্ভাবনা হতে পারে $P(P_1)=1/2$ এবং মহিলা নমুনা এককসমূহের জন্য এই সম্ভাবনা হতে পারে $P(P_2)=1/2$।

এখন একটি নমুনা একক হতে পাওয়া চলকসমূহের মানের ভেক্টর x হলে তা i-তম গণসমষ্টিভুক্ত হবে যদি $P(P_i)\,L_i(x)$ বৃহত্তম হয়। এটিই হলো Bayes নির্ণায়ক পদ্ধতি। এই পদ্ধতির ভিত্তিতে x এর মানের জন্য নির্ণায়ক সাফল্যাঙ্ক (Discriminant score) D হলে নমুনা এককটি i-তম গণসমষ্টিভুক্ত হওয়ার তথ্য-ভিত্তিক (Posterior) সম্ভাবনা হলো

$$P(P_i/D) = \frac{P(D/P_i)\,P(P_i)}{\sum_{i=1}^{k} P(D/P_i)\,P(P_i)}$$

এখানে $P(D/P_i)$ হলো i-তম গণসমষ্টি জানা থাকলে D-এর শর্তাধীন সম্ভাবনা।

উপরিউক্ত সম্ভাবনা $P(P_i/D)$ পাওয়া যায় Bayes নিয়মানুযায়ী। এখন যে কোনো নমুনা এককের জন্য D-এর মান নির্ণয় করা হলে ঐ মানের উপর ভিত্তি করে নমুনা একক ঐ গণসমষ্টিভুক্ত হবে যে গণসমষ্টির জন্য $P(P_i/D)$ বৃহত্তম হবে।

৮.৫ ভুল শ্রেণিভুক্তকরণের সম্ভাবনা (Probability of Misclassification)

সর্বোত্তম সম্ভাব্য (ML) নির্ণায়ক পদ্ধতি অনুসারে যে কোনো একটি নমুনা একক গণসমষ্টি $\mathbf{P_1}$ ভুক্ত হবে, যদি

$$L_1(x) > L_2(x)$$

হয়। এখানে $L_1(x)$ ও $L_2(x)$ হলো যথাক্রমে P_1 ও P_2-এর জন্য সম্ভাব্য ফাংশন [L. F]। যদি $P_1 \sim N_p(\mu_1, \Sigma)$ এবং $P_2 \sim N_p(\mu_2, \Sigma)$ হয়, তাহলে x, P_1 ভুক্ত হবে, যদি $(\mu_1-\mu_2)' \Sigma^{-1} x - \frac{1}{2}(\mu_1-\mu_2)' \Sigma^{-1}(\mu_1+\mu_2) > 0$, অন্যথায় x, P_2 ভুক্ত হবে।

উপরিউক্ত নির্ণায়ক নিয়ম ৮.৪.১ অনুচ্ছেদে আলোচিত হয়েছে। এই নিয়ম হতে ভুল শ্রেণিভুক্তকরণের সম্ভাবনা নির্ণয় করা যায়।

ধরা যাক

$$Y = (\mu_1-\mu_2)' \Sigma^{-1} X$$

হলো রৈখিক নির্ণায়ক চলক এবং Y পরিমিত বিন্যাস অনুসরণ করে। Y-এর দুটি পরামান হলো

$$E(Y) = (\mu_1-\mu_2)' \Sigma^{-1} \mu_1$$

এবং

$$V(Y) = (\mu_1-\mu_2)' \Sigma^{-1} (\mu_1-\mu_2) = \delta$$

এখানে অনুমান করা হচ্ছে যে X i-তম গণসমষ্টিভুক্ত। এই V(Y) হলো দুটি বহু-চলক পরিমিত বিন্যাসের ক্ষেত্রে Mahalanobis দূরত্ব। এখন ভুল শ্রেণিভুক্ত-করণের সম্ভাবনাকে লেখা যায়

$$\begin{aligned}\pi_{21} &= P(x,\ P_2\text{-ভুক্ত}/x,\ P_1\text{-ভুক্ত}) \\ &= P\left[Y \leqslant \tfrac{1}{2}(\mu_1-\mu_2)' \Sigma^{-1}(\mu_1-\mu_2)\right] \\ &= \phi(-\tfrac{1}{2}\delta)\end{aligned}$$

এখানে $\phi(z)$ হলো আদর্শায়িত পরিমিত বিন্যাস ফাংশন। এই বিন্যাস প্রতিসম হওয়ার কারণে এবং শ্রেণিভুক্তকরণ নিয়ম প্রতিসম হওয়ার কারণে লেখা যায়

$$\begin{aligned}\pi_{12} &= P[x\ P_1\text{-ভুক্ত}/x\ P_2\text{-ভুক্ত}] \\ &= \phi(-\tfrac{1}{2}\delta)\end{aligned}$$

আবার ধরা যাক নির্ণায়ক ফাংশন হলো

$$D = x' S^{-1}(\bar{X}_1 - \bar{X}_2) - \tfrac{1}{2}(\bar{X}_1 + \bar{X}_2)' S^{-1}(\bar{X}_1 - \bar{X}_2)$$

এবং x, P_1-ভুক্ত হবে, যদি $D > 0$; অন্যথায় P_2-ভুক্ত হবে। এই নির্ণায়ক ফাংশনের ক্ষেত্রে ভুল শ্রেণিভুক্তকরণের সম্ভাবনা হলো

$$\pi_1 = P(D \leqslant 0) = P(x\ P_2\text{-ভুক্ত}/x\ P_1\text{-ভুক্ত})$$

$$= P\left[\frac{x'b - \mu_1'b}{\sqrt{b'\Sigma b}} \leqslant \frac{\frac{1}{2}(\bar{X}_1 + \bar{X}_2)'b - \mu_1'b}{\sqrt{b'\Sigma b}}\right]$$

$$= \Phi\left[\frac{\frac{1}{2}(\bar{X}_1 + \bar{X}_2)'b - \mu_1'b}{\sqrt{b'\Sigma b}}\right]$$

এখানে $b = S^{-1}(\bar{X}_1 - \bar{X}_2)$। অনুরূপভাবে

$$\pi_2 = \Phi\left[\frac{\mu_2'b - \frac{1}{2}(\bar{X}_1 + \bar{X}_2)'b}{\sqrt{b'\Sigma b}}\right]$$

এই সম্ভাবনা নির্ণয়ের একটি পদ্ধতি Okamoto (1963) আলোচনা করেছেন। Lachenbruch and Mickey (1968) D-এর জটিল বিন্যাস আলোচনা না করেই সম্ভাবনা নির্ণয় পদ্ধতি আলোচনা করেছেন। তাঁদের আলোচিত পদ্ধতি অনুসারে i-তম নমুনার j-তম এককের জন্য পাওয়া যায়

$$D_j = \left\{x_j - \frac{1}{2}\left[\bar{X}_1 + \bar{X}_2 - \frac{1}{n_i - 1}(x_j - \bar{X}_i)\right]\right\}' S_j^{-1}$$

$$\left[\bar{X}_1 - \bar{X}_2 + \frac{(-1)^i}{n_i - 1}(x_j - \bar{X}_i)\right]$$

এখানে

$$S_j^{-1} = \frac{n-3}{n-2}\left[S^{-1} + \frac{C_i}{1 - C_i(x_j - \bar{X}_i)S^{-1}(x_j - \bar{X}_i)} S^{-1}(x_j - \bar{X}_i)(x_j - \bar{X}_i)'S^{-1}\right]$$

$$n = n_1 + n_2,\ C_i = \frac{n_i}{(n_i - 1)(n-2)},\ i = 1, 2$$

এই D_j নির্ণয়ের একটি সহজ সূত্র হলো

$$D_j = \frac{n-3}{n-2}\Big[x_j' S^{-1}(\bar{X}_1 - \bar{X}_2) - \tfrac{1}{2}(\bar{X}_1 + \bar{X}_2)' S^{-1}(\bar{X}_1 - \bar{X}_2)$$

$$+ \frac{1}{1 - C_1(x_j - \bar{X}_1)' S^{-1}(x_j - \bar{X}_1)}$$

$$\times \{ C_1[(x_j - \bar{X}_1)' S^{-1}(x_j - \bar{X}_1)]\,[(x_j - \bar{X}_1)' S^{-1}(\bar{X}_1 - \bar{X}_2)]$$

$$+ C_1[(x_j - \bar{X}_1)' S^{-1}(\bar{X}_1 - \bar{X}_2)]^2$$

$$+ (-1)^1 \frac{2n_1 - 1}{2(n_1 - 1)^2}[(x_j - \bar{X}_1)' S^{-1}(x_j - \bar{X}_1)]^2\}\Big]$$

এখন $D_j > 0$ হলে x_j P_1-ভুক্ত হবে, নতুবা তা P_2-ভুক্ত হবে। তারপর ভুল শ্রেণিভুক্ত হওয়ার সমানুপাত দ্বারা π_1 ও π_2 নিরূপণ করা হয়।

Lachenbruch and Mickey আরো দুটি তথ্যজমান ব্যবহার করে ভুল শ্রেণি-ভুক্ত হওয়ার সম্ভাবনা নির্ণয় করার প্রস্তাব করেছেন। এগুলো হলো

$$\bar{D}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} D_{ij} \qquad s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(D_{ij} - \bar{D}_i)^2, \quad i = 1, 2$$

এই তথ্যজমানের ভিত্তিতে ভুল শ্রেণিভুক্ত হওয়ার সম্ভাবনার নিরূপক হলো

$$\hat{\pi}_1 = \Phi\left(-\frac{\bar{D}_1}{s_1}\right) \text{ এবং } \hat{\pi}_2 = \Phi\left(\frac{\bar{D}_2}{s_2}\right)$$

এখানে $\Phi(z)$ হলো আদর্শায়িত পরিমিত বিন্যাস ফাংশন।

উদাহরণ ৮.২ : উদাহরণ ৮.১-এর উপাত্তের ক্ষেত্রে x_5-এর মানের ভিত্তিতে x_1, x_2, x_3 ও x_4 চলকসমূহের মানকে তিনটি গুণসমষ্টির উপাত্ত বিবেচনা করে নির্ণায়ক বিশ্লেষণ করা যাক।

আলোচিত উপাত্তের ক্ষেত্রে গুচ্ছ গড় ও গুচ্ছ পরিমিত ব্যবধান হলো

গুচ্ছ	গুচ্ছ গড় চলকসমূহ			
	x_1	x_2	x_3	x_4
1	2.632	9.526	3.000	7.474
3	2.250	7.125	2.375	12.500
4	3.125	11.375	3.125	5.500
মোট	2.657	9.400	2.886	8.171
	গুচ্ছ পরিমিত ব্যবধান			
1	1.461	5.450	0.943	1.576
3	0.463	2.416	0.518	2.070
4	0.834	2.669	0.834	2.507
মোট	1.187	4.532	0.867	3.139

এখানে $n_1 = 19$, $n_3 = 8$ এবং $n_4 = 8$। এই তিনটি গুচ্ছের গুচ্ছ-ভেদাঙ্ক সহ-ভেদাঙ্ক ম্যাট্রিক্স হলো

		ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্স			
গুচ্ছ		x_1	x_2	x_3	x_4
1	x_1	2.135	7.316	1.278	0.295
	x_2	7.316	29.708	4.556	0.404
	x_3	1.278	4.556	0.889	0.111
	x_4	0.295	0.404	0.111	2.485
3	x_1	0.214	0.679	0.179	−0.714
	x_2	0.679	5.839	0.661	−3.214
	x_3	0.179	0.661	0.268	−0.500
	x_4	−0.714	−3.214	−0.500	4.286

	x_1	0.696	1.232	0.696	0.929
4	x_2	1.232	7.125	1.232	2.500
	x_3	0.696	1.232	0.696	0.929
	x_4	0.929	2.500	0.929	6.286

এই ভেদাঙ্ক-সহভেদাঙ্ক ম্যাট্রিক্সগুলো সমমাত্রিক। কেননা Box's M-test অনুযায়ী $-2\log\lambda = 17.755$ এবং এটি 28 স্বাধীনতার মাত্রাসহ χ^2-বিন্যাস অনুসরণ করে $[P(\chi^2 \geqslant 17.755) > 0.05]$।

আলোচিত উপাত্তের ক্ষেত্রে চলকসমূহের নির্ণয়ন ক্ষমতা (Discriminating capacity) পর্যালোচনা করা যেতে পারে। প্রতি চলকের গণসমষ্টি গড়সমূহ সমান হলে চলকগুলোর নির্ণয়ন ক্ষমতা ন্যূনতম। চলকসমূহের গণসমষ্টি গড়গুলোর সমতা এক-চলক F-যাচাই তথ্যজমান দ্বারা যাচাই করা যায়। নিচে এই যাচাই তথ্যজমানের ফলাফল দেয়া হলো।

চলক	আন্তঃশ্রেণি বর্গসমষ্টি SS(Between)	বিচ্যুতির বর্গসমষ্টি SS(Within)	F	P-value
x_1	3.090	44.796	1.10	0.344
x_2	72.913	625.487	1.87	0.171
x_3	2.793	22.750	1.96	0.157
x_4	216.235	118.737	29.14	0.000

প্রতি ক্ষেত্রেই F-এর স্বাধীনতার মাত্রা হলো 2 এবং 32। লক্ষ্য করা যাচ্ছে যে চলক x_4-এর মান তিনটি গণসমষ্টির ক্ষেত্রে তাৎপর্যপূর্ণভাবে ভিন্ন। অন্য চলকসমূহের গড়গুলো F-যাচাই তথ্যজমানের দ্বারা তাৎপর্যপূর্ণভাবে ভিন্ন না হলেও চারটি চলকের গড় ভেক্টরগুলো তাৎপর্যপূর্ণভাবে ভিন্ন। এটি Wilk's Λ তথ্যজমান হতে লক্ষণীয়। এখানে $\Lambda = 0.286$, $F = 6.30[P = 0.000]$। এই F-এর স্বাধীনতার মাত্রা হলো 8 এবং 58।

উপরিউক্ত বিশ্লেষণ হতে সিদ্ধান্ত নেয়া যায় যে, আলোচিত উদাহরণের ক্ষেত্রে চলকসমূহের নির্ণয়ন ক্ষমতা আছে। এই চলকসমূহের পুলড অন্তঃগুচ্ছ (Pooled within-group) সংশ্লেষণ ম্যাট্রিক্স হলো

পুলড্ অন্তঃগুচ্ছ সংশ্লেষণ ম্যাট্রিক্স

	x_1	x_2	x_3	x_4
x_1	1.000			
x_2	0.867	1.000		
x_3	0.912	0.798	1.000	
x_4	0.093	0.008	0.096	1.000

উপরিউক্ত উদাহরণের ক্ষেত্রে $k=3$ এবং $P=4$ । সুতরাং $\text{rank}(W^{-1}B) = \min(P, k-1)=2$, এখানে

$$W = W_1 + W_2 + W_3, \quad W_i = \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)'$$

$i = 1, 3, 4$; $n_1 = 19, n_3 = 8, n_4 = 8$; $B = T - W$,

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})(X_{ij} - \bar{X})'$$

X_{ij} হলো i-তম গণসমষ্টির চলকসমূহের ভেক্টরের j-তম মান ; $\bar{X}_i$ এবং $\bar{X}$ হলো যথাক্রমে i-তম গণসমষ্টির এবং সকল নমুনার নমুনা গড় ভেক্টর । যেহেতু $\text{rank}(W^{-1}B) = 2$, সে কারণে আলোচিত উপাত্তের জন্য দুটি নির্ণায়ক ফাংশন পাওয়া যায় । নিচে নির্ণায়ক ফাংশনের সাথে জড়িত ফলাফল দেয়া হলো ।

সারণি ৮.১ : কানুনী নির্ণায়ক ফাংশন এবং এর তাৎপর্য ।

ফাংশন	আই-গেন মান	ভেদাঙ্কের শতকরা হার	ভেদাঙ্কের যোজিত শতকরা হার	কানুনী সংশ্লে-ষাঙ্ক	Λ	χ^2	d.f.	P-value
1	2.082	93.99	93.99	0.822	0.286	38.14	8	0.000
2	0.133	6.01	100.00	0.344	0.882		3	0.282

লক্ষ্য করা যাচ্ছে যে, দুটি নির্ণায়ক ফাংশনের প্রথমটি তাৎপর্যপূর্ণ এবং এটি তিনটি গণসমষ্টির মধ্যে পার্থক্য নির্ণয় করতে $\lambda_1/\Sigma\lambda_j = 0.9399$ বা 93.99% ভেদের ব্যাখ্যা করতে পারে। দ্বিতীয় ফাংশনটি তাৎপর্যপূর্ণ নয় এবং এটি গণসমষ্টির মধ্যে পার্থক্য নির্ণয় করতে $\lambda_2/\Sigma\lambda_j = 0.0601$ বা 6.01% ভেদের ব্যাখ্যা করতে পারে। এখানে আলোচিত ফাংশনের তাৎপর্য যাচাই ৮.৬ অনুচ্ছেদে করা হয়েছে।

আলোচিত নির্ণায়ক ফাংশনদ্বয়ের ক্ষেত্রে কানুনী নির্ণায়ক ফাংশন সহগ (Canonical discriminant function co-efficient) সারণি ৮.২-এ দেয়া হলো।

সারণি ৮.২ ঃ কানুনী নির্ণায়ক ফাংশন সহগ।

চলক	আদর্শায়িত (standardized) ফাংশন		অনাদর্শায়িত (unstandardized) ফাংশন	
	1	2	1	2
x_1	0.343	−2.484	0.290	−2.099
x_2	−0.048	0.103	−0.011	0.023
x_3	−0.609	2.265	−0.722	2.686
x_4	0.961	0.175	0.499	0.091
ধ্রুবক	—	—	−2.663	−3.133

এখানে প্রথম ফাংশনই তাৎপর্যপূর্ণ এবং এই ফাংশনের সহগসমূহ হতে লক্ষ্য করা যাচ্ছে যে পিতার পেশা [x_5] অনুযায়ী জনউর্বরতা (Fertility) সংক্রান্ত তথ্য-সমূহের মধ্যে পার্থক্য নির্দেশ করতে আকাঙ্খিত সন্তানের সংখ্যা (x_3) এবং মায়ের শিক্ষা (x_4) খুবই গুরুত্বপূর্ণ ভূমিকা পালন করেছে। বিষয়টি নির্ণায়ক ফাংশন মানসমূহ (D_1) এবং নির্ণায়ক ফাংশনের জন্য ব্যবহৃত চলকসমূহের মানের সংশ্লেষাঙ্ক হতেও বুঝা যায়। এই সংশ্লেষাঙ্কসমূহ সারণি ৮.৩-এ দেয়া হলো।

সারণি ৮.৩ ঃ নির্ণায়ক বিশ্লেষণে ব্যবহৃত চলক ও নির্ণায়ক ফাংশনের মানের সংশ্লেষাঙ্ক।

চলক	ফাংশন	
	1	2
x_4	0.934*	0.161
x_3	−0.242*	0.098
x_1	−0.164	−0.312*
x_2	−0.229	−0.240*

*-দ্বারা তাৎপর্যপূর্ণ সংশ্লেষাঙ্ক বুঝানো হয়েছে। নির্ণায়ক ফাংশনে চলকের গুরুত্ব অনুসারে সংশ্লেষাঙ্ক উপস্থাপন করা হলো।

লক্ষ্য করা যাচ্ছে যে, প্রথম ফাংশনের ক্ষেত্রে x_4 ও x_3 খুবই গুরুত্বপূর্ণ। আবার, প্রথম ফাংশন ব্যবসা এবং কৃষিকাজ করা পিতাদের ক্ষেত্রে অধিক ভালভাবে পার্থক্য নির্দেশ করে। এটি গ্রুপ সেন্ট্রয়েড ($\bar{D}$) [Group centroid] হতে লক্ষ্য করা যায়। নিচে ফাংশনদ্বয়ের গ্রুপ সেন্ট্রয়েড সারণি ৮.৪-এ দেয়া হলো।

সারণি ৮.৪ ঃ গ্রুপ সেন্ট্রয়েড ($\bar{D}$)।

গ্রুপ	গ্রুপ সেন্ট্রয়েড ($\bar{D}$) ফাংশন	
	1	2
1	−0.439	0.300
3	2.435	−0.177
4	−1.392	−0.536

দ্বিতীয় ফাংশনটি তাৎপর্যপূর্ণ না হলেও, এটি শিক্ষক ($x_5=1$) এবং কৃষিকাজ [$x_5=4$] করা পিতাদের জনউর্বরতা সংক্রান্ত তথ্যের ভিত্তিতে পার্থক্য নির্দেশ করে ভালভাবে। এই পার্থক্য নির্দেশ করার জন্য x_1 ও x_2 খুবই গুরুত্বপূর্ণ।

সারণি ৮.৩ হতে লক্ষ্য করা যাচ্ছে যে, প্রথম ফাংশনের ক্ষেত্রে চলক x_4 নির্ণায়ক ফাংশনের সাথে সবচেয়ে বেশি সংশ্লেষিত। দ্বিতীয় বেশি সংশ্লেষিত চলক হলো x_3। এই শেষোক্ত সংশ্লেষণ ঋণাত্মক চিহ্নযুক্ত। এর অর্থ হলো নির্ণায়ক

ফাংশনের ছোট মান $x_5 = 4$ এর সাথে সম্পর্কিত এবং D_1-এর বড় মান $x_5 = 3$-এর সাথে সম্পর্কিত। এই বিষয়টি সারণি ৮.২-এর আদর্শায়িত সহগ হতেও লক্ষ্য করা যায়।

আবার, সারণি-৮.২ এবং সারণি ৮.৩-এর প্রতি লক্ষ্য করলে দেখা যাবে যে নির্ণায়ক ফাংশনের উপর x_1-এর ধনাত্মক প্রভাব আছে [Blackith and Reyment (1971)]। কিন্তু x_1 প্রথম নির্ণায়ক ফাংশনের সাথে ঋণাত্মকভাবে সংশ্লেষিত (যদিও তাৎপর্যপূর্ণ নয়)। এর কারণ হলো x_4-এর একটি ঋণাত্মক প্রভাব (সাধারণভাবে) x_1-এর উপর আছে এবং x_4 নির্ণায়ক ফাংশনের সাথে বেশি সংশ্লেষিত। এরূপ হওয়ার কারণ হলো নির্ণায়ক ফাংশন সৃষ্টি হওয়ার ক্ষেত্রে x_1 ও x_4 এর প্রভাব সম্পূর্ণ অনপেক্ষ নয়। এ আলোচনা হতে বুঝা যায় যে, চলকসমূহের সংশ্লেষণ নির্ণায়ক ফাংশন সহগকে প্রভাবিত করে থাকে।

উপরিউক্ত বিশ্লেষণের ভিত্তিতে এবং নির্ণায়ক ফাংশনের মানের ভিত্তিতে কোন নমুনা একক কোন শ্রেণিভুক্ত হয়েছে তা দেখানো যায়। এ সম্পর্কে বিশ্লেষিত ফলাফল সারণি ৮.৫-এ উপস্থাপন করা হলো। এখানে লক্ষণীয় বিষয়

সারণি ৮.৫ ঃ নির্ণায়ক সাফলাঙ্ক (Discriminant scores)।

ক্রমিক সংখ্যা	মূল শ্রেণি	নির্ণায়ক বিশ্লেষণের মাধ্যমে সম্ভাব্য শ্রেণি	সম্ভাবনা		দ্বিতীয় সম্ভাব্য শ্রেণি (সম্ভাবনাসহ)		নির্ণায়ক সাফলাঙ্ক
			$P(D/P_i)$	$P(P_i/D)$	শ্রেণি	$P(P_i/D)$	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	1**	4	0.962	0.625	1	0.374	−1.117
							−0.503
2	3	3	0.611	0.999	1	0.001	3.373
							−0.504
3	1	1	0.807	0.538	4	0.461	−1.093
							0.268
4	1	1	0.343	0.472	3	0.438	0.923
							−0.233

5	3	3	0.578	0.999	1	0.001	3.405
							−0.573
6	3	3	0.589	0.999	1	0.001	3.395
							−0.550
7	3	3	0.332	0.468	1	0.445	0.955
							−0.303
8	1	1	0.727	0.680	4	0.319	−0.994
							0.876
9	1	1	0.817	0.533	4	0.466	−1.071
							0.222
10	1	1	0.343	0.472	3	0.438	0.923
							−0.233
11	1	1	0.820	0.531	4	0.468	−1.061
							0.198
12	1**	3	0.327	0.458	1	0.454	0.945
							−0.280
13	3**	1	0.362	0.490	3	0.418	0.901
							−0.186
14	3	3	0.450	0.793	1	0.171	1.377
							−0.866
15	3	3	0.566	0.999	1	0.001	3.416
							−0.597
16	3	3	0.064	0.978	1	0.022	2.662
							2.160
17	4**	1	0.990	0.648	4	0.344	−0.507
							0.172

18	4	4	0.815	0.607	1	0.393	− 1.549
							0.084
19	4	4	0.957	0.629	1	0.370	− 1.095
							− 0.549
20	4	4	0.708	0.767	1	0.231	− 1.118
							− 1.320
21	4	4	0.023	0.979	1	0.021	− 4.068
							− 1.140
22	4	4	0.767	0.761	1	0.238	− 1.162
							− 1.227
23	4	4	0.944	0.635	1	0.363	− 1.062
							− 0.619
24	4**	1	0.991	0.661	4	0.333	− 0.572
							0.312
25	1**	4	0.541	0.658	1	0.335	− 0.587
							− 1.299
26	1	1	0.807	0.538	4	0.461	− 1.093
							0.268
27	1**	4	0.949	0.633	1	0.365	− 1.073
							− 0.596
28	1	1	0.829	0.524	4	0.475	− 1.028
							0.128
29	1	1	0.288	0.521	4	0.330	0.379
							− 1.048
30	1	1	0.205	0.530	3	0.444	1.119
							1.163

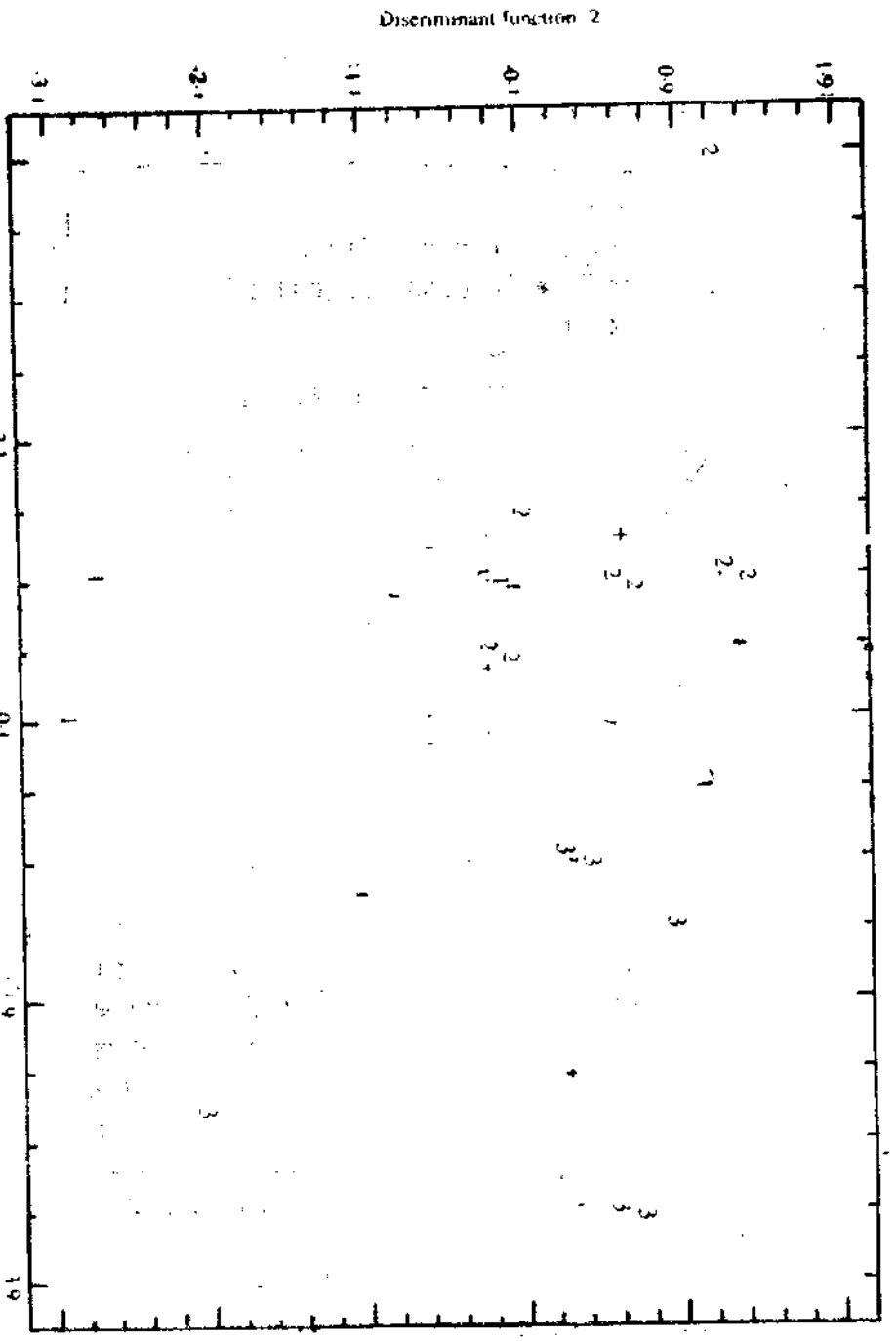
31	1	1	0.616	0.558	4	0.441	−1.354
							0.663
32	1	1	0.695	0.607	4	0.356	−0.053
							−0.461
33	1	1	0.024	0.958	4	0.033	−0.136
							3.011
34	1	1	0.034	0.904	4	0.096	−1.124
							2.807
35	1	1	0.828	0.685	4	0.312	−0.855
							0.754

** নির্ণায়ক বিশ্লেষণের মাধ্যমে ভুল শ্রেণিভুক্ত হিসেবে চিহ্নিত।

হলো যে, 35 এককের মধ্যে একক নির্ণায়ক বিশ্লেষণের মাধ্যমে ভুল শ্রেণিভুক্ত বলে চিহ্নিত। সুতরাং সঠিকভাবে এককসমূহ গুচ্ছভুক্ত হওয়ার শতকরা হার হলো 80.00। এখানে কোন এককের জন্য সম্ভাব্য শ্রেণি চিহ্নিত হয়েছে ঐ এককের জন্য তথ্যভিত্তিক (posterior) সম্ভাবনা নির্ণয়ের মাধ্যমে। উদাহরণ হিসেবে উল্লেখ করা যায় যে, প্রথম এককটি মূলত $x_6 = 1$ শ্রেণিভুক্ত। কিন্তু এটি $x_5 = 4$ শ্রেণিভুক্ত হওয়ার তথ্যভিত্তিক সম্ভাবনা $[P(P_1/D) = 0.625]$ বেশি। সুতরাং প্রথম এককটির $x_5 = 4$ শ্রেণিভুক্ত হওয়ার সম্ভাবনা বেশি। সারণি ৮.৫-এ প্রতিটি এককের জন্যই এরূপ তথ্যভিত্তিক সম্ভাবনা দেয়া আছে। এই সম্ভাবনা Computer program দ্বারা নির্ণয় করা হয়েছে। এই সম্ভাবনার ভিত্তিতে এককসমূহের সম্ভাব্য শ্রেণিভুক্ত হওয়ার সংক্ষিপ্ত ফলাফল সারণি ৮.৬-এ উপস্থাপন করা হলো।

সারণি ৮.৬ ঃ শ্রেণিভুক্তকরণের ফলাফল।

মূল শ্রেণি	একক সংখ্যা	সম্ভাব্য শ্রেণির একক সংখ্যা		
		গুচ্ছ-1	গুচ্ছ-3	গুচ্ছ-4
গুচ্ছ-1	19	15, 78.9%	1, 5.3%	3, 15.8%
গুচ্ছ-3	8	1, 12.5%	7, 87.5%	0, 0.0%
গুচ্ছ-4	8	2, 25.0%	0, 0.0%	6, 75%


Discriminant Analysis for X6
Discriminant function 2
Discriminant function 1
-3.1
-2.6
-2.1
-1.6
-1.1
-0.6
-0.1
0.9
1.9
4.1
3.1
2.1
1.1
0.1
-0.9
-1.9

চিত্র ৮.৩ :

লক্ষ্য করা যাচ্ছে যে, প্রথম শ্রেণির 78.9% একক নির্ণায়ক বিশ্লেষণের মাধ্যমে প্রথম শ্রেণিভুক্ত বলে চিহ্নিত হয়েছে। দ্বিতীয় এবং তৃতীয় শ্রেণির যথাক্রমে 87.5% এবং 75% একক ঐ দুই শ্রেণিভুক্ত বলে চিহ্নিত হয়েছে।

উপরিউক্ত বিশ্লেষণ হতে বুঝা যাচ্ছে যে, কিছু কিছু একক বাস্তবে একটি শ্রেণির অন্তর্ভুক্ত হলেও নির্ণায়ক বিশ্লেষণের মাধ্যমে অন্য শ্রেণিভুক্ত হতে পারে। এক্ষেত্রে এক গুচ্ছের একক অন্য গুচ্ছের সাথে মিশে থাকে। আলোচিত উদাহরণের ক্ষেত্রে এক গুচ্ছের একক কিভাবে অন্য গুচ্ছের সাথে মিশে আছে তা চিত্র ৮.৩-এ উপস্থাপন করা হলো।

৮.৬ নির্ণায়ক ফাংশনের যাচাই (Test of Discriminant Function)

আগেই উল্লেখ করা হয়েছে যে, নির্ণায়ক বিশ্লেষণের ক্ষেত্রে rank $(W^{-1}B) = \min(P, k-1)$ হলে $\min(P, k-1)$ রৈখিক কম্পোজিট বা নির্ণায়ক ফাংশন পাওয়া যায়। কিন্তু এই ফাংশনগুলোর সবগুলোই যে তাৎপর্যপূর্ণ হবে এমন কোনো কথা নেই। যে সকল ফাংশন পরিসংখ্যানিকভাবে তাৎপর্যপূর্ণ ঐগুলোই বিশ্লেষণে রাখা উচিত। এই তাৎপর্য যাচাই করার একটি পদ্ধতি Bartlett (1947) আলোচনা করেছেন। তাঁর প্রস্তাবিত পদ্ধতি হলো $W^{-1}B$ ম্যাট্রিক্স-এর আইগেন মানের তাৎপর্য যাচাই করার পদ্ধতি। এই যাচাই-এর জন্য যাচাই তথ্যজমান হলো

$$V = -\{n-1) - \tfrac{1}{2}(p+k)\} \ln \Lambda$$

এখানে Λ হলো Wilk's Λ এবং

$$\Lambda = \prod_{j=1}^{r} (1+\lambda_j)^{-1}, \quad r = \min(p, k-1)$$

λ_j হলো $W^{-1}B$ এর j-তম আইগেন মান। Bartlett (1947) দেখিয়েছেন যে, এই V-এর বিন্যাস $p(k-1)$ স্বাধীনতার মাত্রাবিশিষ্ট χ^2 বিন্যাস হয়। এখানে নাস্তিকল্পনা হলো, উপাত্তের জন্য k নির্ণায়ক ফাংশন হবে। এখন V-এর মান $p(k-1)$ স্বাধীনতার মাত্রাবিশিষ্ট χ^2-এর বর্জনীয় মান অপেক্ষা বড় হলে নাস্তিকল্পনা বাতিল হবে এবং নির্ণায়ক বিশ্লেষণ তাৎপর্যপূর্ণ বলে বিবেচিত হবে।

উপরিউক্ত যাচাই তথ্যজমানের ন্যায় অনুরূপ যাচাই তথ্যজমান ব্যবহার করে j-তম নির্ণায়ক ফাংশন বা λ_j-এর তাৎপর্য যাচাই করা যায়। সেক্ষেত্রে যাচাই তথ্যজমান হলো

$$V_j = \{(n-1) - \tfrac{1}{2}(p+k)\} \ln(1+\lambda_j)$$

নাস্তিকল্পনার অধীনে এই V_j-এর বিন্যাস হলো $(p+k-2j)$ স্বাধীনতার মাত্রাবিশিষ্ট প্রায় χ^2-বিন্যাস। এই V_j-এর মান $(p+k-2j)$ স্বাধীনতার মাত্রাবিশিষ্ট

χ^2-এর বর্জনীয় মান অপেক্ষা বড় হলে j-তম আইগেন মান তাৎপর্যপূর্ণ হবে। এই V_j-এর মানের ভিত্তিতে কতকগুলো রৈখিক নির্ণায়ক ফাংশন তাৎপর্যপূর্ণ হবে তাও নির্ণয় করা যায়। প্রথম j নির্ণায়ক ফাংশন তাৎপর্যপূর্ণ হবে যদি $[V - V_1 - V_2 - \ldots - V_{j-1}]$ এর মান $(p-j+1)(k-j)$ স্বাধীনতার মাত্রাবিশিষ্ট χ^2-এর বর্জনীয় মান অপেক্ষা বড় হয়।

উদাহরণ ৮.২-এর ক্ষেত্রে

$$V_1 = \{(n-1) - \tfrac{1}{2}(p+k)\} \ln(1+\lambda_1) = 34.33$$

$$V_2 = \{(n-1) - \tfrac{1}{2}(p+k)\} \ln(1+\lambda_2) = 3.81$$

$$V = \{(n-1) - \tfrac{1}{2}(p+k)\} \sum_{j=1}^{r} \ln(1+\lambda_j) = 38.14$$

সুতরাং $V - V_1 = 3.81 < 7.81$ হওয়াতে প্রথম ফাংশনটিই তাৎপর্যপূর্ণ। এখানে 7.81 হলো $(p-j+1)(k-j) = 3$ $[j=2]$ স্বাধীনতার মাত্রাবিশিষ্ট এবং 5% সংশয়মাত্রায় χ^2-এর মান।

উদাহরণ ৮.৩ : উদাহরণ ৮.১-এর ক্ষেত্রে চলক x_6-এর মানের ভিত্তিতে নির্ণায়ক বিশ্লেষণ করা যাক।

আলোচিত উপাত্তকে x_6-এর মানের ভিত্তিতে দুটি শ্রেণিভুক্ত করা যায়, এখানে $n_1 = 12$, $n_2 = 23$। এখানে উভয় শ্রেণির চলক x_1, x_2 এবং x_4 নিয়ে নির্ণায়ক বিশ্লেষণ করা হবে। উপাত্ত থেকে পাওয়া যায়

$$\bar{X}_1 = \begin{bmatrix} 3.167 \\ 11.000 \\ 9.750 \end{bmatrix}, \quad \bar{X}_2 = \begin{bmatrix} 2.391 \\ 8.565 \\ 7.348 \end{bmatrix}$$

$$S_1 = \begin{bmatrix} 0.879 & & \\ 3.455 & 19.273 & \\ -2.227 & -11.273 & 8.568 \end{bmatrix}, \quad S_2 \begin{bmatrix} 1.522 & & \\ 4.860 & 19.984 & \\ -0.324 & -1.933 & 8.874 \end{bmatrix}$$

আলোচিত উপাত্তের ক্ষেত্রে নাস্তিকল্পনা $H_0 : \Sigma_1 = \Sigma_2$ সত্য। কেননা Box's M-test অনুযায়ী $-2\log\lambda = 14.21$। এই $-2\log\lambda$ q স্বাধীনতার মাত্রা-বিশিষ্ট χ^2 – বিন্যাস অনুসরণ করে $[p(\chi^2 \geq 14.21) > 0.05]$। আবার, চলক x_1

x_2 এবং x_4 নির্ণায়ক বিশ্লেষণে অন্তর্ভুক্ত হতে পারে কেননা, Hotelling's T^2 যাচাই তথ্যজমান দ্বারা নাস্তিকল্পনা $H_0 : \mu_1 = \mu_2$ বাতিল হয়। এখানে $T^2 = 0.384$ এবং প্রাসঙ্গিক $F = 3.97$ এবং $P(F \geq 3.97) = 0.017$। এই F-এর স্বাধীনতার মাত্রা হলো 3 এবং 31। Wilk's Λ তথ্যজমান হতেও একই তথ্য পাওয়া যায়। এই উদাহরণের ক্ষেত্রে চলকসমূহের সংশ্লেষণ ম্যাট্রিক্স হলো

অন্তঃগুচ্ছ সংশ্লেষণ ম্যাট্রিক্স

	x_1	x_2	x_4
x_1	1.000		
x_2	0.864	1.000	
x_4	−0.283	−0.383	1.000

আলোচিত উপাত্তের ক্ষেত্রে $p = 3$, $k = 2$। সুতরাং একটি $[\min(p, k-1)]$ নির্ণায়ক ফাংশন পাওয়া যাবে। সারণি ৮.৭-এ নির্ণায়ক বিশ্লেষণের ফলাফল দেয়া হলো।

সারণি ৮.৭ : কানুনী নির্ণায়ক ফাংশনের তাৎপর্য।

ফাংশন	আইগেন মান	ভেদাঙ্কের শতকরা হার,	ভেদাঙ্কের যোজিত শতকরা হার	কানুনী সংশ্লেষাঙ্ক	Λ	χ^2	d.f.	P − value
1	0.384	100.00	100.00	0.527	0.722	10.25	3	0.017

লক্ষ্য করা যাচ্ছে যে নির্ণায়ক ফাংশনটি তাৎপর্যপূর্ণ এবং এটি শ্রেণি দুটির মধ্যে পার্থক্য নির্দেশ করতে ভেদের 100.00 ব্যাখ্যা করতে পারে, এছাড়া ফাংশনটি যে শ্রেণি দুটির মধ্যে ভালভাবে পার্থক্য নির্দেশ করতে পারে তা গ্রুপ সেন্ট্রয়েড হতেও বুঝা যায়। নিচে দুই শ্রেণির উপাত্তের ভিত্তিতে গ্রুপ সেন্ট্রয়েড $(\bar{D})$ উপস্থাপন করা হলো।

	গ্রুপ সেন্ট্রয়েড $(\bar{D})$
শ্রেণি	ফাংশন-1
1	0.833
2	−0.435

এখন প্রশ্ন হলো শ্রেণি দুটির মধ্যে পার্থক্য থাকার জন্য কোন চলকগুলো গুরুত্বপূর্ণ। বিষয়টি কানুনী নির্ণায়ক ফাংশন সহগ সারণি ৮.৮ হতে বুঝা যাবে। লক্ষ্য করা যাচ্ছে যে,

সারণি ৮.৮ ঃ কানুনী নির্ণায়ক ফাংশন সহগ।

চলক	আদর্শায়িত [Standardized]	অনাদর্শায়িত [Unstandardized]
	ফাংশন-1	ফাংশন-1
x_1	0.462	0.404
x_2	0.384	0.086
x_4	0.918	0.310
ধ্রুবক	–	−4.419

মায়ের পেশার ভিত্তিতে জনউর্বরতা সংক্রান্ত তথ্যের মধ্যে পার্থক্য নির্দেশ করতে মায়ের শিক্ষা (x_4) গুরুত্বপূর্ণ ভূমিকা পালন করছে। বিষয়টি নির্ণায়ক সাফলাঙ্ক (D_1) এবং বিশ্লেষণে ব্যবহৃত চলকসমূহ সংশ্লেষাঙ্ক হতেও বুঝা যায়। এই সংশ্লেষাঙ্ক সারণি ৮.৯-এ দেয়া হলো। দেখা যাচ্ছে যে, শ্রেণিদ্বয়ের মধ্যে পার্থক্য নির্দেশ করার জন্য x_4-ই গুরুত্বপূর্ণ ভূমিকা পালন করছে।

সারণি ৮.৯ ঃ নির্ণায়ক বিশ্লেষণে ব্যবহৃত চলক ও নির্ণায়ক ফাংশনের মানের সংশ্লেষাঙ্ক।

চলক	ফাংশন-1
x_4	0.639*
x_1	0.535
x_2	0.432

* সংশ্লেষাঙ্ক তাৎপর্যপূর্ণ

এখন নির্ণায়ক সাফলাঙ্ক (D_1) নির্ণয় করে কোন একক কোন শ্রেণিভুক্ত হয়েছে তা দেখানো যায়। সারণি ৮.১০-এ এই সাফলাঙ্ক দেখানো হলো।

সারণি ৮.১০ ঃ নির্ণায়ক সাফলাঙ্ক (Discriminant score)।

ক্রমিক সংখ্যা	মূল শ্রেণি	নির্ণায়ক বিশ্লেষণের মাধ্যমে সম্ভাব্য শ্রেণি	সম্ভাবনা		দ্বিতীয় সম্ভাব্য শ্রেণি (সম্ভাবনাসহ)		নির্ণায়ক সাফলাঙ্ক
			$P(D/P_i)$	$P(P_i/D)$	শ্রেণি	$P(P_j/D)$	D_i
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	1**	2	0.765	0.605	1	0.395	−0.136
2	1	1	0.558	0.825	2	0.175	1.419
3	1	1	0.928	0.715	2	0.285	0.924
4	1	1	0.939	0.670	2	0.330	0.757
5	1	1	0.744	0.772	2	0.228	1.160
6	1	1	0.680	0.791	2	0.209	1.246
7	1	1	0.737	0.593	2	0.407	0.498
8	1	1	0.586	0.817	2	0.183	1.378
9	1	1	0.934	0.668	2	0.332	0.751
10	1	1	0.939	0.670	2	0.330	0.757
11	1	1	0.866	0.643	2	0.357	0.664
12	1	1	0.803	0.620	2	0.380	0.584
13	2**	1	0.923	0.716	2	0.284	0.930
14	2	2	0.539	0.506	1	0.494	0.180
15	2**	1	0.811	0.752	2	0.248	1.073
16	2**	1	0.618	0.808	2	0.192	1.333
17	2**	1	0.770	0.607	2	0.393	0.542
18	2**	1	0.572	0.522	2	0.478	0.268
19	2	2	0.900	0.656	1	0.344	−0.309

20	2	2	0.350	0.880	1	0.120	−1.369
21	2	2	0.057	0.962	1	0.038	−2.341
22	2	2	0.556	0.825	1	0.175	−1.023
23	2	2	0.893	0.726	1	0.274	−0.569
24	2**	1	0.820	0.744	2	0.251	1.061
25	2	2	0.377	0.873	1	0.127	−1.319
26	2**	1	0.928	0.715	2	0.285	0.924
27	2	2	0.962	0.704	1	0.296	−0.482
28	2**	1	0.668	0.565	2	0.435	0.405
29	2	2	0.996	0.692	1	0.308	−0.440
30	2	2	0.961	0.704	1	0.296	−0.484
31	2	2	0.077	0.955	1	0.045	−2.206
32	2	2	0.690	0.574	1	0.426	−0.036
33	2	2	0.183	0.924	1	0.076	−1.767
34	2	2	0.042	0.967	1	0.033	−2.473
35	2	2	0.144	0.935	1	0.065	−1.896

নির্ণায়ক বিশ্লেষণের মাধ্যমে ভুল শ্রেণিভুক্ত বলে চিহ্নিত।

এই বিশ্লেষণ হতে লক্ষ্য করার বিষয় হলো $D_i > 0$ হলে কোনো একক $x_6 = 1$ শ্রেণিভুক্ত হবে। অন্যথায় তা $x_6 = 2$ শ্রেণিভুক্ত হবে। দেখা যাচ্ছে যে, প্রথম এককটি প্রথম শ্রেণিভুক্ত ছিল। কিন্তু বিশ্লেষণের পরে $D_i < 0$ হওয়াতে এই এককটি দ্বিতীয় শ্রেণিভুক্ত $[x_6 = 2]$ বলে চিহ্নিত। সুতরাং প্রথম এককটি ভুল শ্রেণিভুক্ত। এই বিশ্লেষণে মোট 9 একক ভুল শ্রেণিভুক্ত বলে চিহ্নিত হয়েছে। সুতরাং সঠিক শ্রেণিভুক্ত হওয়ার শতকরা হার হলো 74.29। এখন উভয় শ্রেণির এককের সঠিক শ্রেণিভুক্ত হওয়ার শতকরা হার সারণি ৮.১১-এ উপস্থাপন করা হলো। লক্ষ্য করা

যাচ্ছে যে, প্রথম শ্রেণির 91.7% একক সঠিকভাবে প্রথম শ্রেণিভুক্ত হয়েছে। অপর পক্ষে দ্বিতীয় শ্রেণির 65.2% সঠিকভাবে ঐ শ্রেণিভুক্ত হয়েছে।

সারণি ৮.১১ : শ্রেণিভুক্তকরণের ফলাফল।

মূল শ্রেণি	একক সংখ্যা	সম্ভাব্য শ্রেণির একক সংখ্যা	
		গুচ্ছ-1	গুচ্ছ-2
গুচ্ছ-1	12	11, 91.7%	1, 8.3%
গুচ্ছ-2	23	8, 34.8%	15, 65.2%

নবম অধ্যায়

বহুচলক ভেদাঙ্ক বিশ্লেষণ
(Multivariate Analysis of Variance)

৯.১ সূচনা

ভেদাঙ্ক বিশ্লেষণের ক্ষেত্রে সংগৃহীত উপাত্তের মোট ভেদাঙ্ককে ভেদাঙ্কের পূর্ব নির্ধারিত উৎস অনুসারে বিভাজন করা হয় এবং উৎসসমূহের তাৎপর্য যাচাই করা হয়। একমুখী শ্রেণিবিন্যাসের (one-way classification) ক্ষেত্রে ধরা হয় যে উপাত্ত k গণসমষ্টি হতে চয়ন করা হয়েছে এবং অনুমান করা হয় যে i-তম গণসমষ্টির তথ্যমানসমূহ $N(\mu_i, \sigma^2)$ অনুসরণ করে। এক্ষেত্রে বিশ্লেষণের একটি প্রধান অঙ্গ হলো নাস্তিকল্পনা $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ যাচাই করা। এখানে i-তম গণসমষ্টি হতে চয়ন করা তথ্যমান হলো $x_{i1}, x_{i2}, \cdots, x_{in_i}$ $(i = 1, 2, \cdots, k)$ এবং x_{ij} $(j = 1, 2, \cdots, n_i)$ হলো i-তম গণসমষ্টির j-তম এককের যে কোনো একটি বৈশিষ্ট্যের পরিমাপ। বাস্তবে একাধিক বৈশিষ্ট্যের [p] পরিমাপ করা যায়। ধরা যাক i-তম গণসমষ্টির p বৈশিষ্ট্যের ভিত্তিতে গড় ভেক্টর হলো $\underset{\sim}{\mu}_i$ $(i = 1, 2, \cdots, k)$। তাহলে একচলক (univariate) ভেদাঙ্ক বিশ্লেষণের ন্যায় p-চলক ভেদাঙ্ক বিশ্লেষণের ক্ষেত্রে $H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 = \cdots = \underset{\sim}{\mu}_k$ যাচাই করা যায়। এরূপ নাস্তিকল্পনা যাচাই বা একচলক ভেদাঙ্ক বিশ্লেষণের ন্যায় বহুচলকের জন্য ভেদাঙ্ক বিশ্লেষণ করা হলে তাকে বহুচলক ভেদাঙ্ক বিশ্লেষণ বলা হয়।

ভেদাঙ্ক বিশ্লেষণের ক্ষেত্রে k গণসমষ্টির উপাত্তকে k চর্যার (treatment) উপাত্ত হিসেবে বিবেচনা করা হয়। ধরা যাক i-তম $[i = 1, 2, \dots, k]$ চর্যা j-তম $(j = 1, 2, \dots, n_i)$ প্লটে প্রয়োগ করা হয়েছে এবং ঐ প্লট হতে p চলকের মান পরিমাপ করে X_{ij} নামক একটি $(p \times 1)$ উপাত্ত ভেক্টর পাওয়া গেছে। এই উপাত্ত ভেক্টর যদি একটি উৎস (চর্যা) অনুসারে পাওয়া যায়, তাহলে এই ভেক্টরের জন্য মডেল হলো

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

এখানে μ = উপাত্ত ভেক্টরের সার্বিক প্রভাব বা সার্বিক গড়,

α_i = i-তম চর্যার প্রভাবের ভেক্টর,

ϵ_{ij} = দৈব বিচ্যুতির ভেক্টর এবং এটি $N_p(\sigma, \Sigma)$ অনুসরণ করে।

ধরা যাক $X_{ij} \sim N_p(\mu_i, \Sigma)$ অনুসরণ করে। তাহলে $\mu_i = \mu + \alpha_i$, $i = 1, 2, \cdots, k$ (এখানে ভেক্টরসমূহকে সাধারণ চিহ্নে উপস্থাপন করা হলো)। এক্ষেত্রে বিশ্লেষণের একটি মুখ্য উদ্দেশ্য হলো নাস্তিকল্পনা

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

যাচাই করা বা

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k$$

যাচাই করা। এখানে

$$\alpha_i = [\alpha_{i1} \ \alpha_{i2} \cdots \alpha_{ip}]'$$

৯.২ বহুচলক ভেদাঙ্ক বিশ্লেষণের জন্য অনুমান (Assumption for Multivariate Analysis of Variance)

একচলক ভেদাঙ্ক বিশ্লেষণের জন্য দুটি অনুমান যেমন করতে হয়, বহুচলক ভেদাঙ্ক বিশ্লেষণের জন্যও তেমনি দুটি অনুমান করতে হয়। এই অনুমান দুটি হলো :

(i) বিশ্লেষণের জন্য উপাত্তসমূহ দৈব পদ্ধতিতে চয়ন করা হয় এবং

(ii) উপাত্ত $N_p(\mu_i, \Sigma)$ হতে চয়ন করা হয়। এই শেষোক্ত অনুমানের দ্বারা বুঝা যায় যে প্রতিটি গণসমষ্টি একই ভেদাঙ্কবিশিষ্ট পরিমিত বিন্যাস অনুসরণ করে।

আলোচিত অনুমান বহাল থাকে কিনা তা পরীক্ষা করার জন্য p চলকের প্রতিটির জন্য ভিন্ন ভিন্নভাবে অনুমান বহাল আছে কিনা পর্যালোচনা করা যায়। অবশ্য প্রতি চলকের ক্ষেত্রে অনুমান বহাল থাকলেই যে চলকসমূহের যুগ্ম বিন্যাস পরিমিত হবে তা বলা যায় না। চলকসমূহের পরিমিত বিন্যাস সম্পর্কে সিদ্ধান্ত নেয়ার জন্য Andrews et al. (1973)-এর প্রস্তাবিত পদ্ধতি প্রয়োগ করা যায়। Box's (1949) M-যাচাই পদ্ধতি প্রয়োগ করে গণসমষ্টিগুলোর একই ভেদাঙ্ক ম্যাট্রিক্স সম্পর্কে সিদ্ধান্ত নেয়া যায়।

বহুচলক ভেদাঙ্ক বিশ্লেষণ করার আগে আরেকটি বিষয়ে লক্ষ্য রাখতে হয়। তাহলো চলকসমূহ সংশ্লেষিত কিনা? চলকগুলো সংশ্লেষিত না হলে বহুচলক বিশ্লেষণ হতে অর্থবহ তথ্য পাওয়া যাবে না। কাজেই বিশ্লেষণের শুরুতেই নাস্তিকল্পনা $H_0 : P = I$ যাচাই করা দরকার। এখানে P হলো গণসমষ্টি সংশ্লেষাঙ্ক ম্যাট্রিক্স এবং I হলো আইডেনটিটি ম্যাট্রিক্স। Bartlett (1947) এই নাস্তিকল্পনা যাচাই পদ্ধতি আলোচনা করেছেন। এখানে ৫.৫.৩-সূত্রে এই নাস্তিকল্পনা যাচাই-এর জন্য যাচাই তথ্যজমান দেয়া আছে।

৯.৩ একমুখী শ্রেণিবিন্যাস (The One-way Classification)

ধরা যাক একটি নিরীক্ষায় k চর্যা আছে এবং প্রতিটি চর্যার ফলাফল হতে p চলকের মান পরিমাপ করা যায়। ধরা যাক i-তম $(i=1, 2, \ldots, k)$ চর্যার j-তম প্লট হতে পরিমাপ করা চলকসমূহের ভেক্টর হলো $X_{ij}(P\times 1)$। তাহলে

$$X_{ij}=\mu+\alpha_i+\epsilon_{ij}$$

$$i=1, 2, \cdots, k; \quad j=1, 2, \ldots, n_i,$$

$$\epsilon_{ij}\sim N_p(0, \Sigma)$$

এই প্রতিকৃতির ক্ষেত্রে যাচাই করতে হবে নাস্তিকল্পনা

$$H_0: \mu_1=\mu_2=\ldots=\mu_k$$

এবং বিকল্প কল্পনা হলো

$$H_A: \mu_i\neq\mu_{i'} \quad (i\neq i')$$

উক্ত নাস্তিকল্পনা যাচাই করার জন্য সম্ভাব্য অনুপাত নির্দেশক (likelihood ratio criterion) হলো

$$\Lambda=|W|/|T|$$

এখানে

$$W_i=\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)(X_{ij}-\bar{X}_i)'$$

$$W=W_1+W_2+\ldots+W_k$$

$$T=W+B$$

$$B=\sum_{i}^{k} n_i(\bar{X}_i-\bar{\bar{X}})(\bar{X}_i-\bar{\bar{X}})'$$

$$\bar{X}_i=n_i^{-1}\sum_{j=1}^{n_i}X_{ij}, \quad \bar{\bar{X}}=n^{-1}\sum_{i}^{k}\sum_{j}^{n_i}X_{ij}, \quad n=\sum_{i=1}^{k}n_i$$

নাস্তিকল্পনার অধীনে $W\sim W_p(\Sigma, n-k)$, $B\sim W_p(\Sigma, k-1)$ এবং এরা অনপেক্ষ। কাজেই $n\geqslant p+k$ হলে

$$\Lambda=|W|/|W+B|\sim\Lambda(p, n-k, k-1)$$

এখানে Λ হলো Wilk's Λ এবং

$$\Lambda = \prod_{j=1}^{p} (1+\lambda_j)^{-1}$$

$\lambda_1, \lambda_2, \ldots, \lambda_p$ হলো $W^{-1}B$-এর আইগেন মানসমূহ। Λ-এর ছোট মানের জন্য নাস্তিকল্পনা বাতিল হয়। আবার Λ-কে F-বিন্যাসে পরিণত করা যায়, যেখানে

$$\frac{(n-k+p+1)\{1-\sqrt{\Lambda}(p, n-k, k-1)\}}{p\sqrt{\Lambda}(p, n-k, k-1)}$$

$$\sim F_{p(k-1),\,(k-1)(n-k-p+1)}$$

চর্যাসমূহের প্রভাবের সমতা যাচাই $W^{-1}B$-এর বৃহত্তম আইগেন ব্যবহার করেও করা যায়। ধরা যাক λ_r হলো বৃহত্তম আইগেন মান। তাহলে এই λ_r প্রয়োগ করে তথ্যজমান θ_r নির্ণয় করা যায়, যেখানে $\theta_r = \lambda_r/(1+\lambda_r)$। নাস্তিকল্পনার অধীনে θ_r এর বিন্যাসের পরামানসমূহ হলো

$$S = \min(p, k-1),\ m = \frac{1}{2}[|k-p-1|-1],\ n = \frac{1}{2}[n-k-p-1]$$

এখন $\theta_r \leqslant \theta_{\alpha;\,s,\,m,\,n}$

হলে নাস্তিকল্পনার বিপক্ষে কোনো যুক্তি নেই বলে বিবেচিত হবে। এখানে $\theta_{\alpha;\,s,\,m,\,n}$ হলো θ_r এর বিন্যাসের উচ্চ $100\alpha\%$ মান। Heck (1960), Pillai and Bantegui (1959) and Pillai (1964, 1965, 1967) λ_r এর বিন্যাস এবং তার $100\alpha\%$ মান নির্ণয় করেছেন।

উদাহরণ ৯.১ : উদাহরণ ৮.১-এর ক্ষেত্রে $x_6 = 1$ এর উপাত্তসমূহকে একটি চর্যার উপাত্ত এবং $x_6 = 2$ এর উপাত্তসমূহকে অন্য একটি চর্যার উপাত্ত বিবেচনা করে এবং প্রতি প্লট হতে চলক x_1, x_2 ও x_4 পরিমাপ করা হয়েছে বিবেচনা করে একমুখী বহুচলক ভেদাঙ্ক বিশ্লেষণ করা যাক।

আলোচিত উপাত্তের ক্ষেত্রে

$$T = \begin{bmatrix} 47.886 & & \\ 159.800 & 698.400 & \\ -16.943 & -120.400 & 334.971 \end{bmatrix}$$

$$W = \begin{bmatrix} 43.145 & & \\ 144.913 & 651.652 & \\ -31.630 & -166.522 & 289.467 \end{bmatrix}$$

$$B = \begin{bmatrix} 4.741 & & \\ 14.887 & 46.748 & \\ 14.687 & 46.122 & 45.504 \end{bmatrix}$$

এখানে $p=3$, $k=2$, $n_1=12$, $n_2=23$, $n=35$ এবং $W^{-1}B$ ম্যাট্রিক্স-এর আইগেন মান হলো $\lambda=0.384$। চর্যা দুটির উপাত্তকে দুটি গণসমষ্টির উপাত্ত বিবেচনা করা হলে উক্ত গণসমষ্টিদ্বয়ের ক্ষেত্রে $\Sigma_1=\Sigma_2$ [উদাহরণ ৮.৩ দ্রষ্টব্য]। উদাহরণ ৮.৩ হতে আরো লক্ষ্য করা যায় যে $|R| = 0.214$। সুতরাং H_0 : P = I যাচাই করার জন্য যাচাই তথ্যজমান হলো $-2\log\lambda=53.96$ এই $-2\log\lambda$ এর বিন্যাস হলো 3 স্বাধীনতার মাত্রাবিশিষ্ট χ^2 বিন্যাস। সুতরাং x_1, x_2 ও x_4 সংশ্লেষিত বিবেচনা করা যায়।

এখন নাস্তিকল্পনা $H_0 : \mu_1=\mu_2$ যাচাই করার জন্য প্রাসঙ্গিক $\Lambda=0.7223$ এবং এর প্রাসঙ্গিক $F=3.97$। এই F-এর স্বাধীনতার মাত্রা হলো 3 এবং 31 এবং $p(F \geqslant 3.97)=0.017$। সুতরাং চর্যাদ্বয়ের প্রাসঙ্গিক গড় ভেক্টরদ্বয়ে পার্থক্য নেই নাস্তিকল্পনা বাতিল করা যায়।

MANOVA TABLE

ভেদাঙ্কের উৎস	d.f.	SS	Wilk's Λ
চর্যা	1	B	
বিচ্যুতি	33	W	$\Lambda = \dfrac{\lvert W \rvert}{\lvert W+B \rvert} = 0.7223$
মোট	34	T	

গড় ভেক্টরে সমতা নেই অর্থাৎ মায়ের পেশা (x_6) পরিবর্তনের সাথে জীবিত জন্মগ্রহণ করা সন্তানের সংখ্যা (x_1), বিবাহিত জীবনকাল (x_2) এবং মায়ের

শিক্ষার স্তরে (x_4) পরিবর্তন লক্ষ্য করা যায়। এখানে লক্ষণীয় বিষয় হলো x_6 এর পরিবর্তনের সাথে x_1, x_2 ও x_4 এর প্রতিটির মধ্যে পরিবর্তন তাৎপর্যপূর্ণ কিনা। এ ব্যাপারে একচলক F-যাচাই প্রয়োগ করা যায়। সারণি ৯.১-এ এই F-যাচাই-এর ফলাফল দেয়া হলো।

সারণি ৯.১ : একচলক F-যাচাই এর ফলাফল।

চলক	SS(Between)	SS(Within)	F	P-value
x_1	4.741	43.145	3.63	0.066
x_2	46.748	651.652	2.37	0.133
x_4	45.504	289.467	5.19	0.029

লক্ষ্য করা যাচ্ছে যে মায়ের পেশার পরিবর্তন মূলত শিক্ষার স্তরের পরিবর্তনের কারণে তাৎপর্যপূর্ণ।

৯.৪ কনট্রাস্ট-এর যাচাই (Test of Contrast)

একমুখী শ্রেণিবিন্যাসের ক্ষেত্রে অনুমান করা হয়েছে

$$X_{ij} \sim N_p(\mu_i, \Sigma) \quad (i = 1, 2, \cdots, k ; \; j = 1, 2, \ldots, n_i)$$

এক্ষেত্রে বিশ্লেষণের মুখ্য উদ্দেশ্য হলো

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

যাচাই করা। আলোচিত k গণসমষ্টির গড়ের ভিত্তিতে ধরা যাক একটি কনট্রাস্ট হলো

$$\sum_{i=1}^{k} a_i \mu_i \quad (i = 1, 2, \cdots, k)$$

অনেক সময় গড়সমূহের সমতা যাচাই না করে নাস্তিকল্পটি

$$H_0 : \sum_{i=1}^{k} a_i \mu_i = 0$$

যাচাই করতে হয়। এক্ষেত্রে বিকল্প কল্পনা হলো

$$H_1 : \mu_i \neq \mu_j \quad (i \neq j = 1, 2, \cdots, k)$$

এই নাস্তিকল্পনার জন্য যাচাই তথ্যজমান হলো

$$\Lambda(p, n-k, 1) \sim |W| / |W+C|$$

এখানে

$$C=\left(\sum_{i=1}^{k} a_i \bar{X}_i\right)\left(\sum_{i} a_i \bar{X}_i'\right)\Big/\left(\sum_{i=1}^{k} \frac{a_i^2}{n_i}\right)$$

এখানে a_i-গুলো হলো এমন জানা মান যেন $\Sigma a_i = 0$ হয়।

বাস্তবে C ম্যাট্রিক্স পাওয়া যায়

$$n\overset{*}{\Sigma} = W+C$$

হতে [Mardia et al (1988)]।

৯.৫ দ্বিমুখী শ্রেণিবিন্যাস (Two-way Classification)

ধরা যাক একটি উপাত্ত ম্যাট্রিক্সের তথ্যমানসমূহকে A ও B নামক দুটি উপাদানের স্তর অনুসারে ভাগ করা যায়, যেখানে উপাদানদ্বয়ের স্তর হলো যথাক্রমে r এবং c। মনে করা যাক A ও B-এর যে কোনো স্তরের জন্য n অনপেক্ষ উপাত্ত আছে এবং ঐগুলো n একক হতে সংগ্রহ করা। আরো মনে করা যাক যে প্রতিটি একক হতে p চলকের মান পরিমাপ করা হয়েছে। এখন প্রাপ্ত উপাত্তের জন্য একটি মডেল হলো

$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

$$i = 1, 2, \cdots, r\ ;\ j = 1, 2, \cdots, c\ ;\ k = 1, 2, \cdots, n$$

এখানে X_{ijk} হলো $(p \times 1)$ উপাত্ত ভেক্টর,

μ = উপাত্তের সার্বিক প্রভাব,

α_i = উপাদান A-এর i-তম স্তরের প্রভাব,

β_j = উপাদান B-এর j-তম স্তরের প্রভাব,

$(\alpha\beta)_{ij}$ = A-এর i-তম স্তরের সাথে B-এর j-তম স্তরের মিশ্র প্রভাব,

ϵ_{ijk} = দৈব বিচ্যুতির $(p \times 1)$ ভেক্টর।

এখানে অনুমান করা যাক যে $\epsilon_{ijk} \sim N_p(0, \Sigma)$ এবং i, j এবং k-এর সকল মানের জন্য ϵ_{ijk} অনপেক্ষ। আরো অনুমান করা যাক যে, A-এর i-তম স্তরের প্রাসঙ্গিক B-এর j-তম স্তর হতে n তথ্যমান সংগৃহীত হয়েছে। এই অনুমানের সুবিধা হলো যে, মোট বর্গসমষ্টি ম্যাট্রিক্সকে সুচারুভাবে ভেদাঙ্কের উৎস অনুসারে বিভিন্ন বর্গসমষ্টি ম্যাট্রিক্সে বিভাজন করা যায়। এখানে মোট বর্গসমষ্টি ম্যাট্রিক্স হলো

$$T = \Sigma\Sigma\Sigma(X_{ijk} - \bar{X})(X_{ijk} - \bar{X})'$$

এই T-কে বিভাজন করে লেখা যায়

$$T = R + C + I + W$$

যেখানে

$$R = cn\,\Sigma(\bar{X}_{i..} - \bar{X})(\bar{X}_{i} - \bar{X}_{...})'$$

$$C = rn\,\Sigma(\bar{X}_{.j.} - \bar{X})(\bar{X}_{.j.} - \bar{X})'$$

$$I = n\Sigma\Sigma(\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X})(\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X})'$$

$$W = \Sigma\Sigma\Sigma(X_{ijk} - \bar{X}_{ij.})(X_{ijk} - \bar{X}_{ij.})'$$

$$= \Sigma\Sigma A_{ij}$$

এখানে $$A_{ij} = \sum_{k} (X_{ijk} - \bar{X}_{ij.})(X_{ijk} - \bar{X}_{ij.})'$$

$$\bar{X} = \frac{1}{rcn}\Sigma\Sigma\Sigma X_{ijk}, \quad \bar{X}_{i..} = \frac{1}{cn}\sum_{j}\sum_{k} X_{ijk}$$

$$\bar{X}_{.j.} = \frac{1}{rn}\sum_{i}\sum_{k} X_{ijk}, \quad \bar{X}_{ij.} = \frac{1}{n}\sum_{k} X_{ijk}$$

এই বিশ্লেষণের মুখ্য উদ্দেশ্য হলো

$$H_0 : \alpha_i = 0$$

$$H_0 : \beta_j = 0$$

$$H_0 : (\alpha\beta)_{ij} = 0$$

নাস্তিকল্পনাসমূহ যাচাই করা। এই নাস্তিকল্পনাসমূহের অধীনে $T \sim W_p(\Sigma, rcn-1)$। আবার, নাস্তিকল্পনা সত্য হোক না হোক $A_{ij} \sim W_p(\Sigma, n-1)$ এবং A_{ij} সমূহ অনপেক্ষভাবে বিন্যাসিত। কাজেই $W \sim W_p(\Sigma, rc(n-1))$। এছাড়া $H_0 : \alpha_i = 0$ এবং $H_0 : \beta_j = 0$ সত্য হোক বা না হোক, সকল $(\alpha\beta)_{ij}$ সমান হলে, $I \sim W_p(\Sigma, (r-1)(c-1))$, আবার $R \sim W_p(\Sigma, r-1)$ এবং $C \sim W_p(\Sigma, c-1)$ এবং R, C, I ও W অনপেক্ষভাবে বিন্যাসিত। কাজেই $H_0 : (\alpha\beta)_{ij} = 0$ যাচাই করার জন্য যাচাই তথ্যজমান হলো

$$\Lambda = |W| / |W+I| \sim \Lambda(P, rc(n-1), (r-1)(c-1))$$

এখানে Λ-এর মান ছোট হলে নাস্তিকল্পনা বাতিল বলে পরিগণিত হবে। অবশ্য Λ-কে F বিন্যাসেও পরিণত করা যায়।

এই নাস্তিকল্পনা যাচাই করার জন্য $I(W+I)^{-1}$-এর বৃহত্তম আইগেন মানের উপরও নির্ভর করা যায়। এই আইগেন মানের বিন্যাস হলো $\theta(P, \nu_1, \nu_2)$। এই বিন্যাসের শতকরা মানগুলো Foster and Rees (1957) নির্ণয় করেছেন।

এখানে $\nu_1 = rc(n-1)$, $\nu_2 = (r-1)(c-1)$

কোনো পরীক্ষায় $n = 1$ হলে W-এর স্বাধীনতার মাত্রা হয় শূন্য। সে কারণে $n = 1$ হলে $H_0 : (\alpha\beta)_{ij} = 0$ নাস্তিকল্পনা যাচাই করা যায় না।

$H_0 : \beta_j = 0$ সত্য হলে $C \sim W_p(\Sigma, c-1)$ এখন α_i এবং $(\alpha\beta)_{ij}$ গুলো সমান হোক বা না হোক β_j গুলো সমান কিনা তা যাচাই করার জন্য যাচাই তথ্যজমান হলো

$$\Lambda = |W| / |W+C| \sim \Lambda(P, rc(n-1), c-1)$$

এক্ষেত্রেও Λ-এর ছোট মানের জন্য নাস্তিকল্পনা বাতিল বলে পরিগণিত হয়। এই নাস্তিকল্পনাও $C(W+C)^{-1}$-এর বৃহত্তম আইগেন মানের বিন্যাস প্রয়োগ করে যাচাই করা যায়। ধরা যাক θ_1 হলো $C(W+C)^{-1}$-এর বৃহত্তম আইগেন মান। এই θ_1-এর বিন্যাস হলো

$$\theta(P, rc(n-1), c-1)$$

অনুরূপভাবে α_i গুলো সমান যাচাই করার জন্য যাচাই তথ্যজমান হলো

$$\Lambda = |W| / |W+R| \sim \Lambda(P, rc(n-1), r-1)$$

এই যাচাই তথ্যজমান β_j-এর মানের উপর নির্ভর করে না।

উপরে $H_0 : (\alpha\beta)_{ij} = 0$ যাচাই করার পদ্ধতি আলোচনা করা হয়েছে। এই যাচাই-এর মাধ্যমে শাস্তিকল্পনা সত্য হতে পারে বা কোনো উপাত্তে মিশ্র প্রভাব নাও থাকতে পারে [$n = 1$ হলে]। সেক্ষেত্রে বিচ্যুতি ম্যাট্রিক্স W এর পরিবর্তে ম্যাট্রিক্স E নির্ণয় করতে হয়, যেখানে

$$E = \Sigma\Sigma\Sigma (X_{ijk} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...})(X_{ijk} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...})'$$
$$= I + W$$

এক্ষেত্রে β_j-এর তাৎপর্য যাচাই করার জন্য যাচাই তথ্যজমান হবে

$$|E| / |E + C| \sim \Lambda(P, rcn - r - c + 1, c - 1)$$

অবশ্য এই যাচাই তথ্যজমানের পরিবর্তে $C(E+C)^{-1}$-এর বৃহত্তম আইগেন মান-কেও যাচাই-এর জন্য ব্যবহার করা যায়। যদি বৃহত্তম আইগেন মান θ_3 হয়, তাহলে θ_3 এর বিন্যাস হবে

$$\theta(P, rcn - r - c + 1, c - 1)$$

উদাহরণ ৯.২ : উদাহরণ ৮.১-এর উপাত্তের ক্ষেত্রে x_1, x_2 ও x_4 এর মান-সমূহকে x_5 ও x_6 চলকের মানের ভিত্তিতে শ্রেণিবিন্যাস করে দ্বিমুখী ভেদাঙ্ক বিশ্লেষণ করা যাক।

ধরা যাক চলক x_5-এর মানের ভিত্তিতে উপাত্তসমূহ স্তম্ভে বিভক্ত এবং x_6-এর মানের ভিত্তিতে ঐগুলো সারিতে বিভক্ত। তাহলে সারি ও স্তম্ভের কোষসমূহে এককের সংখ্যা হবে নিম্নরূপ।

সারণি ৯.২ : বিভিন্ন শ্রেণিভুক্ত এককের সংখ্যা।

x_6/x_5	1	3	4	মোট
1	8	4	—	12
2	11	4	8	23
মোট	19	8	8	35

ধরা যাক আলোচিত উপাত্তের জন্য x_6-এর জন্য বর্গসমষ্টি ম্যাট্রিক্স হলো R, x_5-এর জন্য বর্গসমষ্টি ম্যাট্রিক্স হলো C, x_5 ও x_6-এর মিশ্র প্রভাবের জন্য তা I এবং বিচ্যুতির বর্গসমষ্টি ম্যাট্রিক্স হলো W। তাহলে,

$$R=\begin{bmatrix} 4.113 & & \\ 12.420 & 37.506 & \\ 5.093 & 15.381 & 6.308 \end{bmatrix},\quad C=\begin{bmatrix} 5.794 & & \\ 24.320 & 106.196 & \\ -27.677 & -129.901 & 178.052 \end{bmatrix}$$

$$I=\begin{bmatrix} 6.820 & & \\ 29.117 & 123.937 & \\ 7.766 & 22.870 & 39.736 \end{bmatrix},\quad W=\begin{bmatrix} 31.159 & & \\ 93.943 & 430.761 & \\ -2.125 & -28.750 & 110.875 \end{bmatrix}$$

$$T=\begin{bmatrix} 47.886 & & \\ 159.800 & 698.400 & \\ -16.943 & -120.400 & 334.971 \end{bmatrix}$$

এখন x_5 ও x_6 এর মিশ্রপ্রভাব যাচাই করার জন্য যাচাই তথ্যজমান হলো

$$\Lambda = |W|/|W+1| = 0.836$$

এই Λ-এর প্রাসঙ্গিক $F=1.835$। এর স্বাধীনতার মাত্রা 3 এবং 28। কিন্তু $P(F \geq 1.835)=0.164$ হওয়াতে মা-বাবার পেশার মিশ্র প্রভাব উপাত্তের ভেদ পর্যালোচনায় কোনো প্রভাব ফেলতে পারেনি। এক্ষেত্রে পিতার পেশার এবং মাতার পেশার প্রভাবের তাৎপর্য যাচাই করার জন্য যাচাই তথ্যজমান হবে, যথাক্রমে

$$\Lambda = |E|/|E+R| \sim \Lambda(p, rcn-r-c+1, r-1)$$

এবং $$\Lambda = |E|/|E+C| \sim \Lambda(p, rcn-r-c+1, c-1)$$

এখানে $$E=I+W=\begin{bmatrix} 37.979 & & \\ 123.060 & 554.698 & \\ 5.641 & -5.880 & 150.611 \end{bmatrix}$$

আলোচিত উপাত্তের ক্ষেত্রে x_5 ও x_6 এর মিশ্র প্রভাব তাৎপর্যহীন হলেও এক-চলক বিশ্লেষণের মাধ্যমে লক্ষ্য করা যাচ্ছে যে দম্পতির বিবাহিত জীবনকালে

(x_2) স্বামী-স্ত্রীর পেশার $[x_5$ ও $x_6]$ একটি তাৎপর্যপূর্ণ প্রভাব আছে। সারণি ৯.৩-এ এই একচলক বিশ্লেষণের ফলাফল দেয়া হলো।

সারণি ৯.৩ ঃ মিশ্রপ্রভাব যাচাই-এর জন্য একচলক বিশ্লেষিত ফলাফল।

চলক	$SS(\alpha\beta)_{ij}$	SS(বিচ্যুতি)	F	P-value
x_1	4.113	31.159	3.96	0.056
x_2	82.427	430.761	14.36	0.023
x_3	0.022	110.875	3.70	0.939

সকল ক্ষেত্রেই F-এর স্বাধীনতার মাত্রা হলো 1 এবং 30।

সন্তান উৎপাদন সংক্রান্ত উপাত্তের $[x_1, x_2, x_4]$ উপর স্বামীর পেশার (x_5) কোনো তাৎপর্যপূর্ণ প্রভাব আছে কিনা তা যাচাই করার জন্য নাস্তিকল্পনা H_0 : $\beta_j = 0$ যাচাই করতে হবে। এই নাস্তিকল্পনার জন্য যাচাই তথ্যজমান হলো

$$\Lambda = |E| / |E + C| = 0.407 \sim \Lambda(3, 31, 2)$$

এই Λ-এর প্রাসঙ্গিক $F = 5.29$ এবং এর স্বাধীনতার মাত্রা হলো 6 এবং 56 এবং $P(F \geq 5.29) = 0.000$ হওয়াতে সন্তান উৎপাদন সংক্রান্ত উপাত্তের উপর স্বামীর পেশার উচ্চ তাৎপর্যপূর্ণ প্রভাব আছে সিদ্ধান্ত নেয়া যায়।

এই তাৎপর্যপূর্ণ প্রভাবের কারণ হলো x_2 এবং x_4 কে x_5 এর মান অনুসারে শ্রেণিভুক্ত করলে বিভিন্ন শ্রেণির মধ্যে তাৎপর্যপূর্ণ পার্থক্য পরিলক্ষিত হয়। বিষয়টি একচলক বিশ্লেষণ হতে লক্ষ্য করা যায়। একচলক বিশ্লেষণের ফলাফল সারণি ৯.৪-এ দেয়া হলো।

সারণি ৯.৪ ঃ x_5 এর প্রভাব যাচাই করার জন্য একচলক F তথ্যজমান।

চলক	$SS(\hat{\beta}_j)$	SS (বিচ্যুতি)	F	P-value
x_1	5.794	31.159	2.789	0.077
x_2	106.196	430.761	3.698	0.037
x_4	178.052	110.875	24.088	0.000

এখানে সকল ক্ষেত্রেই F-এর স্বাধীনতার মাত্রা হলো 2 এবং 30।

আলোচিত উপাত্তের উপর স্ত্রীর পেশার (x_6) প্রভাবের তাৎপর্য যাচাই ($H_0 : \alpha_1' = 0$) করার জন্য যাচাই তথ্যজমান হলো

$$\Lambda = |E| / |E+R| = 0.876 \sim \Lambda(3, 31, 1)$$

এবং প্রাসঙ্গিক $F = 1.32$ $(P > 0.05)$ । কাজেই সন্তান উৎপাদন সংক্রান্ত উপাত্তের উপর স্ত্রীর পেশার (x_6) কোনো তাৎপর্যপূর্ণ প্রভাব নেই। বিষয়টি একচলক বিশ্লেষণ সারণি ৯.৫ হতেও লক্ষ্য করা যায়।

সারণি ৯.৫ ঃ x_6 এর প্রভাব যাচাই করার জন্য একচলক F-তথ্যজমান।

চলক	SS(α_1)	SS (বিচ্যুতি)	F	P-value
x_1	4.113	31.159	3.96	0.056
x_2	37.506	430.761	2.612	0.117
x_4	6.308	110.875	1.707	0.201

সকল ক্ষেত্রেই F-এর স্বাধীনতার মাত্রা 1 এবং 30 ।

সারণি ৯.৬ ঃ MANOVA TABLE

ভেদাঙ্কের উৎস	d.f.	SS	Wilk's Λ
স্ত্রীর পেশা	1	R	$0.876 \sim \Lambda(3, 31, 1)$
স্বামীর পেশা	2	C	$0.407 \sim \Lambda(3, 31, 2)$
স্ত্রীর পেশা × স্বামীর পেশা	1	I	$0.836 \sim \Lambda(3, 28, 1)$
বিচ্যুতি	30	W	—
মোট	34	T	

এখানে বহুচলক ভেদাঙ্ক বিশ্লেষণ সম্পর্কে একটি ধারণা ব্যক্ত করা হলো। এ সম্পর্কে বিস্তারিত জানার জন্য Mardia et al (1988), Bock (1975) আলোচনা করা যেতে পারে।

সহায়ক গ্রন্থপঞ্জি

1. Anderson, E. (1954) : Efficient and inefficient methods of measuring specific differences, In Statistics and Mathematics in Biology, O. Kampthorne (Ed), Ames : I owa State College Press, 98-107.

2. Anderson, E. (1957) : A semi-graphical method for the analysis of complex problems, Technometrics, 2, 387-392.

3. Anderson, E. (1960) : Some Stochastic process models for intelligence test scores. In Mathematical Methods in the Social Sciences, K. J. Arrow et al (eds), Stanford C. A., Stanford University Press, 205-220.

4. Anderson, T. W. (1963) : Asymptotic theory for principal component analysis, Ann. Math. Statist. , 34, 122-148.

5. Andrews, D. F. (1972) : Plots of high dimensional data, Biometrics, 28, 125-136.

6. Andrews, D. F., Gnanadesikan, R. and Warner, J. L. (1973) : Methods for assessing multivariate normality. In Multivariate Analysis III. ed. P. R. Krishnaiah, Academic Press, New York.

7. Arnold, S. J. (1979) : A test for clusters, Journal of Marketing Research, 16, 545-551.

8. Ball, G. H. (1971) : Classification Analysis, Stanford Research Institute, SRI Project 5533.

9. Bartlett, M. S. (1947) : Multivariate Analysis, Jour. Roy. Statist. Soc. B. , 9, 176-197.

10. Bartlette, M. S. (1951) : The effect of standardization on a χ^2 approximation in factor analysis, Biometrika, 38, 337-344.

11. Bartlett, M. S. (1954) : A note on multiplying factors for various chisquared approximations, Jour. Roy. Statist. Soc. B, 16, 296-298.

12. Beale, E.M.L. (1970) : Selecting an optimum subset. In Integer and Non-linear Programming (Abadie, J. ed), North-Holland, Amsterdam.

13. Beale, E. M: L., Kendall, M.G. and Mann, D.W. (1967) : The discarding of variables in multivariate analysis, Biometrika, 54, 357-366.

14. Bhuyan, K.C. (1984) : Principal component regression to study production function of boro rice. J. U. Review, Part-1.

15. Bhuyan, K. C. (1995) : Socio-economic factors influencing child mortality in Bangladesh-A case study, Journal of Family Welfare, 41(1), 15-23.

16. Bhuyan, K.C. and Nair, G.A. (1995) : Structural changes in the relationships of body dimensions of four species of pulmonate slugs. Biometrical Journal, 37(8), 1005-1015.

17. Blackith, R.E. and Reyment, R.A. (1971) : Multivariate Morphometrics, Academic Press, New York.

18. Bock, R. D. (1975) : Multivariate Statistical Methods in Behavioural Research, McGraw-Hill, New York.

19. Box, G. E. P. (1949) : A general distribution theory for a class of likelihood criteria, Biometrika, 36, 317-346.

20. Box, G. E. P. (1950) : Problems in the analysis of growth and wear curves, Biometrics, 6, 362-389.

21. Cattell, R. B. (1952) : Factor Analysis, Harper, New York.

22. Cattell, R. B. (1966) : The scree test for the number of factors, Multivariate Behavioural Research, 1, 245-276.

23. Cattell, R. B. and Jespars, J. (1967) : A general plas mode (No. 30-10-5-2) for factor analytic excercises and research. Multivariate Behavioural Research Monographs, 67-3, 1-212.

24. Cormack, R. M. (1971) : A review of classification, Jour. Roy. Statist. Soc. A., 134, 321-353.

25. Dagnelie, P. (1975) : Analyse Statistique a Plusieurs Variables, Vander, Brussels.

26. Dillon, W.R. and Goldstein, M. (1984) : Multivariate Analysis : Methods and Applications, John Wiley and Sons, Inc., New York.

27. Everitt, B. (1974) : Cluster Analysis, Heineman Educational, London.

28. Everitt, B. S. (1978) : Graphical Techniques for Multivariate Data, North-Holland, New York.

29. Fienberg, S. E. (1979) : Graphical Methods in Statistics. The American Statistician, 33, 165-178.

30. Foster, F.G. and Rees, D.H. (1957) : Upper percentage points of the generalized beta distribution, Biometrika, 44, 237-247.

31. Friedman, H.P. & Rubin, J. (1967) : On some invariant criteria for grouping data, Jour. Amer. Stats. Assoc., 62, 1159-1178.

32. Gaffar, A. F. (1996) : Socio-economic Factors Influencing Child Mortality in North-Eastern Libya. Unpublished M. Sc. Thesis, Garyounis University, Libya.

33. Gamati, Y. M. (1992) : A Comprehensive Statistical Study of Multivariate Data of Milk Frequency Trials Conducted in Ghot Sultan Poultry and Diary Project. Unpublished M. Sc. Thesis, Garyounis University, Libya.

34. Gnanadesikan, R. (1977) : Methods for Statistical Data Analysis of Multivariate Observations, John Wiley and Sons, Inc., New York.

35. Gleason, T. C. (1976) : On redundancy in canonical analysis, Psychological Bulletin, 83(6), 1004-1006.

36. Gower, J. C. and Ross, G. J. S. (1969) : Minimum spanning trees and single linkage cluster analysis, Applied Statistics, 18, 54-64.

37. Greenstreet, R. L. and Connor, R. J. (1974) : Power of tests for equality of covariance matrices, Techometrics, 16, 27-30.

38. Harman, H. H. (1976) : Modern Factor Analysis (3rd edition), Chicago University Press.

39. Hartigan, J. A. (1973) : Minimum mutation fits to a given tree, Biometrics. 29, 53-65.

40. Hartigan, J. A. (1975) : Clustering Algorithoms, John Wiley, New York.

41. Horst, P. (1965) : Factor Analysis of Data Matrices, Hold : Rinehart and Winston, New York.

42. Horn, J. L. (1965) : A rationale and test for the number of factors in factor analysis, Psychometrika, 30, 179.

43. Huff, D. L. and Black, W. (1978) : A multivariate graphical display for regional analysis. In Graphical Representation of Multivariate Data, Peter, C. C. Wang (ed). 199-218, Academic Press, New York.

44. Jardine, N. and Sibson, R. (1971) : Mathematical Taxonomy, John Wiley and Sons, Inc. , New York.

45. Johnson, S. C. (1967) : Hierarchical clustering schemes. Psychometrika, 32, 241-254.

46. Johnson, R. M. (1972) : How can you tell if things are really clustered ? Market Facts, Inc.

47. Jolliffe, I. T. (1972) : Discarding variables in principal component analysis, 1 : artificial data, Appl. Stats. , 21, 160-173.

48. Jolliffee, I. T. (1973). Discarding variables in principal component analysis, II : real data, Appl. Stats., 22, 21-31.

49. Joreskog, K. G. and Lawley, D. N. (1968) : New methods in maximum likelihood factor analysis British Jour. Math. Stats., Psychology, 21, 85-96.

50. Kaiser, H. F. (1958) : The varimax criterion for analytic rotation in factor analysis, Psychometrika, 23, 187-200.

51. Kendall, M. G. (1957) : A course in Multivariate Analysis. Griffins, Statistical Monographs and Courses.

52. Kshirsagar, A. M. (1972) : Multivariate Analysis, Marcell Dekkar, New York.

53. Lachenbruch, P. A. (1967) : An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis, Biometrics, 23, 639-645.

54. Lachenbruch, P. A. (1975) : Discriminant Analysis, Hafner Press, New York.

55. Lachenbruch, P. A. Sneeringer, C. and Revo, L. T. (1973) : Robustness of the linear and quadratic discriminant function to certain types of non-normality, Communications in Statistics 1, 39-57.

56. Lawley, D. N. (1940) : The estimation of factor loadings by the method of maximum likelihood, Proceedings of the Roy. Statist. Soc. of Edinburgh, 60, 64-82.

57. Lawley, D. N. (1942) : Further investigations in factor estimation. Proceedings of the Roy. Statist. Soc. of Edinburgh Section-A, 61, 176-185.

58. Lawley, D. N. (1943) : The application of the maximum likelihood method to factor analysis, British Journal of Psychology, 33, 172-175,

59. Lawley, D.N. (1956) : Tests of significance for the latent roots of covariance and correlation matrices, Biometrika, 43, 128-136.

60. Lawley, D. N. and Maxwell, A. E. (1971) : Factor Analysis in a Statistical Methods. American Elsevier, New York.

61. Lee, K. L. (1979) : Multivariate tests for clusters. Jour. Amer. Stats. Assoc., 74, 708 – 714.

62. Mahaghub, El-Amin, A. (1996) Testing Regression Equality to study the Fertility Differentials in North-Eastern Libya, M. S. C. Thesis, Garyounis University, Libya.

63. Massy, W. F. (1965) : Principal component analysis in exploratory data research. Jour. Amer. Stats. Assoc., 60, 234 — 256.

64. Mardia, K. V., Kent, J. T. and Bibby, J. M. (1988) : Multivariate Analysis, Academic Press, London.

65. Mc Clain, J. O. and Rao. V. R. (1975) : CLUSTSIZ : A program to test for the equality of clustering of a set of objects Jour. Marketing Research, 12, 456 — 460.

66. Mac Naughton-Smith, P. (1965) : Some Statistical and Other Numerical Techniques for Classifying Individuals, Home Office Research Report No. 6, London, H. M. S. O.

67. Maxwell, A. E. (1967) : Calculating maximum likelihood factor loadings, Jour. Roy. Statist. Soc., A. 127, 238 — 241.

68. Miller, J. K. (1975) : In defense of the general canonical correlation index : Reply to Niewander and Wood, Psychological Bulletin, 82. 207 209.

69. Mittelhammer, R. C. and Baritelle, J. L. (1977) : On two strategies for choosing principal components in regression analysis, Amer. Jour. Agril., Econ., 59, 336 — 343.

70. Pearson, E. S. and Hartley, H. O. (1972) : Biometrika Tables for statisticians, Vol. 2. Cambridge University Press, Cambridge.

71. Rao, C. R. (1955) : Estimation and tests of significance in factor analysis, Psychometrika, 20, 93 — 111.

72. Rummel, R. J. (1970) : Applied Factor Analysis, Evanston, J. L., North-Western University Press.

73. Sneath, P. H. A. (1957) : The application of computer taxonomy, Jour. General Microbiology, 17, 201 — 226.

74. Spearman, C. (1904) : The proof and measurement of association between two things, Amer. Jour. Psychol., 15, 72 and 202.

75. Stewart, D. K. and Love, W. A. (1968) : A general canonical correlation index, Psychological Bulletin, 70, 160 — 163.

76. Tatsuoka, M. M. (1971) : Multivariate Analysis, John Wiley and Sons, Inc., New York.

77. Thompson, M. L. (1978) : Selection of variables in multiple regression, I, II, Int. Statist. Rev., 46, 1 – 20, 129 – 146.

78. Ward, J. (1963) : Hierarchical grouping to optimize an objective function, Jour. Amer. Stats. Assoc., 58, 236 – 244.

79. Wahl, P. W. and Kronmal, R. A. (1977) : Discriminant functions when covariances are unequal and sample sizes are moderate, Biometrics, 33, 479 – 484.

80. Wilks, S. S. (1949) : Sample criteria for testing equality of means, equality of variances and equality of covariances in a normal multivariate distribution, Ann. Math. Statist., 17, 257 – 281.

81. Wolfe, J. H. (1970) : Pattern clustering by multivariate mixture analysis. Multivariate Baha. Res. 5, 329 – 350.

82. Wollenberg, A. L. Vanden (1977) : Redundancy analysis : An alternative for canonical correlation analysis, Psychometrika, 41, 207 – 219.

83. Yao, Y. (1965) : An approximate degrees of freedom solution to the multivariate Behrens-Fisher problem, Biometrika, 52, 139 – 147.