

Lecture Notes in Statistics

Kaizhao Liu

May 25, 2025

Contents

1	Introduction	5
1.1	In the Heartbeat of Our Real World	5
1.2	The Enchanting Tapestry of Interdisciplinary Insights	5
1.3	Organization	6
2	Three Basic Objectives of Statistics	7
3	The Curse of Dimensionality	9
4		11
5	Mathematical Statistics	13
5.1	Populations, Samples, and Models	13
5.1.1	Populations and samples	13
5.1.2	Parametric and nonparametric models	13
5.2	Statistics, Sufficiency, and Completeness	14
5.2.1	Statistics	14
5.2.2	Sufficiency and minimal sufficiency	14
5.2.3	Completeness	14
5.3	Statistical Decision Theory	15
5.3.1	Decision rules, loss functions, and risks	15
5.3.2	Admissibility and optimality	16
5.4	Statistical Inference	16
5.5	Asymptotic Criteria and Inference	16
5.5.1	Consistency	16
5.5.2	Asymptotic Bias, Variance, and MSE	16
5.5.3	Asymptotic Inference	17
5.6	Unbiased Estimation	17
5.7	Estimation in Parametric Models	17
5.8	Hypothesis tests	17
5.9	Confidence sets	17
6	Multivariate Data Analysis	19
6.1	The Multivariate Normal Distributions	19
6.1.1	Asymptotic Distributions of Sample Means and Covariance Matrices	19
6.1.2	The Noncentral χ^2 and F Distributions	19
6.1.3	Quadratic Forms	20
6.1.4	Spherical and Elliptical Distributions	20
6.2	Multivariate Integration	20
6.2.1	Exterior Products	20
6.2.2	The Multivariate Gamma Function	21
6.2.3	Miscellaneous Jacobians	21
6.2.4	Invariant Measures	21
7	von Mises Expansion	23
7.1	Int	23
7.2	Influence Function	23
7.3	Infinitesimal Jackknife	23
7.4	Perturbing a Training Input	24

I	Causal Inference	25
8	Counterfactuals	27
8.1	Binary Treatment	27
8.2	28
9	CPDAG	29
9.1	PDAGs	29
9.2	MECs	29
9.3	A Transformational Characterization of MECs	30
9.4	Characterization of CPDAGs	30
9.5	Algorithm:	32
9.6	Algorithm:	32
9.6.1	Compelled and Reversible Edge Insertions	32
9.6.2	The InsertU Operator	32
9.6.3	The DeleteU Operator	32
9.6.4	The InsertD Operator	32
9.6.5	The DeleteD Operator	33
9.6.6	The ReverseD Operator	33
9.6.7	The MakeV Operator	33
9.7	Algorithm: Reversible MCMC on MECs	33
9.7.1	The InsertU Operator	33
9.7.2	The DeleteU Operator	33
9.7.3	The InsertD Operator	33
9.7.4	The DeleteD Operator	33
9.7.5	The RemoveV Operator	33
9.8	Counting Sizes of MECs	33
9.8.1	Size of MEC Determined by the Number of Vertices	33

Chapter 1

Introduction

1.1 In the Heartbeat of Our Real World

Statistics is a discipline where we harness the power of numbers to understand data. Its relevance and applications are deeply intertwined with real-world scenarios. Therefore, a solid statistical analysis must first prove its merit in addressing real-world challenges before we delve into its mathematical underpinnings. This mirrors the approach in computational mathematics and probability theory.

However, this might appear starkly different from what you've heard about pure mathematics. Pure math often delves into areas like algebra, number theory, and combinatorics, or it might explore more philosophical realms like mathematical logic, or even bridge with theoretical physics through geometry and analysis. But at their core, these are all frameworks for modeling the world quantitatively. Take combinatorics, for instance: its essence would be lost if we were merely counting systems without origin or purpose. Similarly, what would be the point of studying geometry or topology if they didn't in some way reflect the world around us? A classic example is the advent of analysis. It began as a means to describe object motion but has since permeated countless fields. Even disciplines like number theory, which at first glance seem esoteric, have their roots in deeper philosophical pursuits, aiming for comprehensive understanding.

In essence, while pure mathematicians might appear to be delving into the profound, they're fundamentally exploring challenges that, albeit sometimes abstract, have ties to our tangible world. This interconnectedness between various fields and the real world underscores the essence of our research. After all, the real world remains the ultimate benchmark for the validity and relevance of our academic endeavors.

1.2 The Enchanting Tapestry of Interdisciplinary Insights

Statistics, inherently applied, finds its inspirations from the needs across diverse fields. This includes the sciences – spanning computational science, physics, chemistry, biology, medicine, psychology, and geography – as well as the liberal arts, encompassing economics, finance, sociology, semantics, aesthetics, and numerous others I haven't detailed here. Consequently, our research is deeply interdisciplinary. We value the influence of these diverse disciplines, incorporating their challenges and ideas, which enrich and elevate the realm of statistics. We keep with the pace of them. Moreover, we recognize and harness profound mathematical advancements, arming ourselves with sophisticated tools and methods to address these multifaceted problems.

For instance, one of the early impacts of computer science on statistics was the development of bootstrap resampling. Prior to the computational revolution, statisticians heavily relied on stringent assumptions and theoretical derivations to make population inferences. These theoretical methods often imposed conditions that might not align with real-world data, such as assuming normal distributions or large sample sizes. However, with the advent of more powerful computers and advancements in computer science, Bradley Efron introduced the method of **bootstrap resampling** in the late 1970s. This innovative approach allows statisticians to resample or draw repeated samples from observed data, enabling the estimation of variability and confidence intervals for statistics without necessitating strict theoretical assumptions.

As another illustration, the influence of neuroscience on statistics is notably exemplified by the introduction of neural networks. Traditional statistical methods often make assumptions about data, such as linearity, independence, or homoscedasticity. While these methods have been robust

and enduring, they can face challenges when confronted with complex, non-linear data structures prevalent in today's vast datasets. Neural networks, designed to mimic the intricate workings of the human brain's neurons, exhibit the capacity to model and capture intricate, non-linear relationships in data.

1.3 Organization

Let us restate, **the goal of statistics** is to deal with data encountered in our real world. According to this goal, this lecture note focuses on the interplay between real world applications and theoretical development. As this subject is applied in nature, many of our theoretical investigations are motivated by demand in various related areas, such as sciences like computational science, physics, chemistry, biology, psychology, or liberal arts like

Chapter 2

Three Basic Objectives of Statistics

Chapter 3

The Curse of Dimensionality

Parallel to the development of the early twentieth century statistics, a new branch of science emerges, namely what we called computer science now. They were first used for calculating bullets. The theory of computational mathematics also revives as this new computational tool appeared. Numerical methods for calculus, linear algebra, and differential equations are extensively studied and implemented in the computers.

Chapter 4

Chapter 5

Mathematical Statistics

5.1 Populations, Samples, and Models

A typical statistical problem can be described as follows:

1. (**planning experiments**) One or a series of random experiments is performed;
2. (**collecting data**) some data from the experiments are collected;
3. (**statistical analysis of the data**) extract information from the data, interpret the results, and draw some conclusions

Statistical analysis of the data may contain:

- **descriptive data analysis:** such as the mean, median, range, standard deviation, etc., and some graphical displays, such as the histogram and box-and-whisker diagram, etc..
- **statistical inference**
- **decision theory**

We focus on more sophisticated methods of analyzing data, that is, the last two listed above.

5.1.1 Populations and samples

We formalize our statistics theory in the language of modern probability. First we need to specify some terms.

Definition 5.1.1 (data set). The data set is viewed as a **realization/observation** of a random element defined on a probability space (Ω, \mathcal{F}, P) related to the random experiment.

Definition 5.1.2 (population). The probability measure P is called the population.

Definition 5.1.3 (sample). The random element that produces the data is called a **sample** from P .

Definition 5.1.4 (known). A population P is known if and only if $P(A)$ is known for every event $A \in \mathcal{F}$.

5.1.2 Parametric and nonparametric models

Definition 5.1.5 (statistical model). A statistical model is a set of assumptions on the population P in a given problem.

Remark 5.1.6. It is often postulated to make the analysis possible or easy. Postulated models are often based on knowledge of the problem under consideration, and testing the correctness of them is also a task of statistical inference and decision theory.

Example 5.1.7 (parametric model). The population P is in a given parametric family \mathcal{P} .

Example 5.1.8 (nonparametric model). The population P is in a given nonparametric family \mathcal{P} .

Example 5.1.9 (semi-parametric model). A nonparametric model having a parametric component.

5.2 Statistics, Sufficiency, and Completeness

5.2.1 Statistics

Definition 5.2.1 (statistic). A measurable function $T(X)$ of X is called a statistic if T is a known function.

From a probabilistic point of view, the information within the statistic $T(X)$ is contained in the σ -field $\sigma(T(X))$. (This is because if two statistics generate the same σ -field, then there exist a bijection between them.) From this viewpoint, values of a statistic is non-material.

Any $T(X)$ is a random element. If the distribution of X is unknown, then the distribution of $T(X)$ may also be unknown although T is a known function. Finding the form of the distribution of $T(X)$ is one of the major problems in statistical inference and decision theory.

5.2.2 Sufficiency and minimal sufficiency

We now seek to describe a fully informative statistic $T(X)$. The concept definitioned below describes what we know by fully informative.

Definition 5.2.2 (sufficiency). Let X be a sample from an unknown population $P \in \mathcal{P}$, where \mathcal{P} is a family of populations. A statistics $T(X)$ is said to be sufficient for $P \in \mathcal{P}$ if and only if the conditional distribution of X given T is known (does not depend on P)

Remark 5.2.3. The concept of sufficiency depends on the given family \mathcal{P} . If T is sufficient for $P \in \mathcal{P}$, then T is also sufficient for $P \in \mathcal{P}_0 \subset \mathcal{P}$.

This definition can be interpreted as follows. Once we observe X and compute a sufficient statistic $T(X)$, the original data X do not contain any further information concerning the unknown population P and can be discarded.

Finding a sufficient statistic in general involves a 'guess and verify' procedure. However, for families of populations having p.d.f's, a simple way of finding sufficient statistics is to use the factorization theorem.

Theorem 5.2.4 (The factorization theorem). *Suppose that X is a sample from $P \in \mathcal{P}$ and \mathcal{P} is a family of probability measures on $(\mathbb{R}^n, \mathcal{B}^n)$ dominated by a σ -finite measure ν . Then $T(X)$ is sufficient for $P \in \mathcal{P}$ if and only if there are nonnegative Borel functions h (known) on $(\mathbb{R}^n, \mathcal{B}^n)$ and g_P (unknown) on the range of T such that*

$$\frac{dP}{d\nu} = g_P(T(x))h(x)$$

For the proof of this theorem, we need a lemma concerning a family of probability measures.

Lemma 5.2.5. *If a family \mathcal{P} is dominated by a σ -finite measure, then \mathcal{P} is dominated by a probability measure $Q = \sum_{i=1}^{\infty} c_i P_i$, where c_i 's are nonnegative constants with $\sum_{i=1}^{\infty} c_i = 1$ and $P_i \in \mathcal{P}$.*

Actually, there are many sufficient statistics for a given family \mathcal{P} , as the following theorem shows.

Theorem 5.2.6. *If T is a sufficient statistics and $T = \phi(S)$, where ϕ is measurable and S is another statistics, then S is sufficient.*

Remark 5.2.7. Note that $\sigma(T) \subset \sigma(S)$ so T provides a further reduction of the data than S .

Therefore a question is raised. Is there a sufficient statistics that provides maximal reduction of the data?

Definition 5.2.8 (minimal sufficiency). Let T be a sufficient statistic for $P \in \mathcal{P}$. T is called a minimal sufficient statistic if and only if, for any other statistic S sufficient for $P \in \mathcal{P}$, there is a measurable function ϕ such that $T = \phi(S)$ a.s. \mathcal{P} .

5.2.3 Completeness

Definition 5.2.9 (ancillary). A statistic $V(X)$ is said to be ancillary if its distribution does not depend on the population P .

Definition 5.2.10 (first-order ancillary). A statistic $V(X)$ is said to be first-order ancillary if $E[V(X)]$ is independent of P .

Recalling the definition of a complete orthogonal system in functional analysis, we define the completeness of a statistic similarly. For a statistic $T(X)$, the measurable map $T : (\Omega, \mathcal{F}) \rightarrow (\mathcal{T}, \mathcal{F}_T)$ induces a family of distributions $\mathcal{P}^T := \{P \circ T^{-1}\}_{P \in \mathcal{P}}$ on $(\mathcal{T}, \mathcal{F}_T)$. For $g : (\mathcal{T}, \mathcal{F}_T) \rightarrow (\mathbb{R}, \mathcal{B})$ a \mathcal{P}^T integrable Borel function, denote $\langle P, g \rangle_T := E_P[g(T(X))] = \int_{\mathcal{T}} g dP^T = P^T(g)$.

Definition 5.2.11 (completeness). A statistic $T(X)$ is said to be complete for \mathcal{P} if and only if for any Borel function g with $\langle P, g \rangle_T = 0, \forall P \in \mathcal{P}$, we have $g = 0$ a.s. \mathcal{P}^T .

Definition 5.2.12 (boundedly completeness). A statistic $T(X)$ is said to be boundedly complete for \mathcal{P} if and only if for any bounded Borel function g with $\langle P, g \rangle_T = 0, \forall P \in \mathcal{P}$, we have $g = 0$ a.s. \mathcal{P}^T .

Remark 5.2.13. If T is (boundedly) complete and $S = \phi(T)$ for a measurable ϕ , then S is (boundedly) complete.

5.3 Statistical Decision Theory

5.3.1 Decision rules, loss functions, and risks

Let X be a sample from a population $P \in \mathcal{P}$.

Definition 5.3.1 (statistical decision). A statistical decision is an **action** that one take after one observe X .

Definition 5.3.2 (action space). Denote the set of allowable actions by \mathbb{A} . Let $\mathcal{F}_{\mathbb{A}}$ be a σ -field on \mathbb{A} . Then the measurable space $(\mathbb{A}, \mathcal{F}_{\mathbb{A}})$ is called the action space.

Definition 5.3.3 (decision rule). Let \mathbf{X} be the range of X and $\mathcal{F}_{\mathbf{X}}$ be a σ -field on \mathbf{X} . A decision rule is a known measurable function (a statistic) T from $(\mathbf{X}, \mathcal{F}_{\mathbf{X}})$ to $(\mathbb{A}, \mathcal{F}_{\mathbb{A}})$.

Definition 5.3.4 (randomized decision rule). A randomized decision rule is a function δ on $\mathbf{X} \times \mathcal{F}_{\mathbb{A}}$ s.t., for every $A \in \mathcal{F}_{\mathbb{A}}$ $\delta(\cdot, A)$ is a Borel function and, for every $x \in \mathbf{X}$ $\delta(x, \cdot)$ is a probability measure on $(\mathbb{A}, \mathcal{F}_{\mathbb{A}})$.

Remark 5.3.5. An alternate way to describe a randomized decision rule is to specify the method of **simulating** the action from \mathbb{A} for each $x \in \mathbf{X}$.

The construction or selection of decision rules cannot be done without any criterion about the performance of decision rules. In statistical decision theory, we set a criterion using a loss function.

Definition 5.3.6 (loss function). A loss function L is a function from $\mathcal{P} \times \mathbb{A}$ to $[0, \infty)$ and is Borel on $(\mathbb{A}, \mathcal{F}_{\mathbb{A}})$ for each fixed $P \in \mathcal{P}$.

Remark 5.3.7. The loss function for a randomized rule δ is definitionined as

$$L(P, \delta, x) = \int_{\mathbb{A}} L(P, a) d\delta(x, a)$$

A decision rule with small loss is preferred, but it is difficult to compare loss functions since they are random. For this reason, we introduce the risk function.

Definition 5.3.8 (risk). The risk of a decision rule T , which is the average loss for T , is definitionined to be

$$R_T(P) = E[L(P, T(X))] = \int_{\mathbf{X}} L(P, T(x)) dP_X(x)$$

Remark 5.3.9. The risk of a randomized rule δ is definitionined as

$$R_{\delta}(P) = E[L(P, \delta, X)] = \int_{\mathbf{X}} \int_{\mathbb{A}} L(P, a) d\delta(x, a) dP_X(x)$$

Definition 5.3.10 (as good as). A rule T_1 is as good as another rule T_2 if and only if

$$R_{T_1}(P) \leq R_{T_2}(P) \quad \forall P \in \mathcal{P}$$

Remark 5.3.11. If the inequality is strict for at least one $P \in \mathcal{P}$, we say T_1 is **better** than T_2 .

Definition 5.3.12 (equivalent). Two decision rules T_1 and T_2 are equivalent if and only if

$$R_{T_1}(P) = R_{T_2}(P) \quad \forall P \in \mathcal{P}$$

Definition 5.3.13 (\mathfrak{F} -optimal). Let \mathfrak{F} be a class of decision rules. If there is a decision rule T_* that is as good as any other rule in \mathfrak{F} , then T_* is said to be \mathfrak{F} -optimal.

5.3.2 Admissibility and optimality

Consider a given decision problem with a given loss $L(P, a)$.

Definition 5.3.14 (admissibility). Let \mathfrak{F} be a class of decision rules. A decision rule $T \in \mathfrak{F}$ is called \mathfrak{F} -admissible if and only if there does not exist any $S \in \mathfrak{F}$ that is better than T .

In view of the fact that an optimal rule often does not exist, statisticians adopt the following two approaches to choose a decision rule.

The first approach is to define a class \mathfrak{F} of decision rules that have some desirable properties and then try to find the best rule in \mathfrak{F} .

Definition 5.3.15 (bias). In an estimation problem, the bias of an estimator $T(X)$ of a real-valued parameter ϑ of the unknown population is defined to be $b_T(P) = E[T(X)] - \vartheta$. An estimator $T(X)$ is said to be unbiased for ϑ if and only if $b_T(P) = 0$ for any $P \in \mathcal{P}$.

Definition 5.3.16 (invariance).

The second approach is to consider some characteristic R_T of $R_T(P)$ for a given decision rule T , and then minimize R_T over $T \in \mathfrak{F}$.

5.4 Statistical Inference

In statistical inference, we make an inference about the unknown population based on the sample X and **inference procedure**, without using any loss function, although any inference procedure can be cast in decision-theoretic terms as a decision rule.

There are three main types of inference procedures:

- **point estimators**
- **hypothesis tests**
- **confidence sets**

We will discuss them in the following sections.

5.5 Asymptotic Criteria and Inference

5.5.1 Consistency

Definition 5.5.1 (consistency of point estimators). Let $X = (X_1, \dots, X_n)$ be a sample from $P \in \mathcal{P}$ and $T_n(X)$ be a point estimator of ϕ for every n .

- (i) $T_n(X)$ is called consistent for ϕ if and only if $T_n(X) \rightarrow_p \phi$ w.r.t any $P \in \mathcal{P}$.
- (ii) Let $\{a_n\}$ be a sequence of positive constants diverging to ∞ . $T_n(X)$ is called a_n -consistent for ϕ if and only if $a_n[T_n(X) - \phi] = O_p(1)$ w.r.t. any $P \in \mathcal{P}$.
- (iii) $T_n(X)$ is called strongly consistent for ϕ if and only if $T_n(X) \rightarrow \phi$ a.s. w.r.t. any $P \in \mathcal{P}$.
- (iv) $T_n(X)$ is called L_r -consistent for ϕ if and only if $T_n(X) \rightarrow_{L_r} \phi$ w.r.t any $P \in \mathcal{P}$ for some fixed $r > 0$.

Remark 5.5.2. Consistency is actually a concept relating to a sequence of estimators.

5.5.2 Asymptotic Bias, Variance, and MSE

Definition 5.5.3 (asymptotic expectation). Let ξ, ξ_1, ξ_2, \dots be random variables and $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \rightarrow \infty$ or $a_n \rightarrow a > 0$. If $a_n \xi_n \rightarrow_d \xi$ and $E|\xi| < \infty$, then $\frac{E\xi}{a_n}$ is called an asymptotic expectation of ξ_n .

Lemma 5.5.4.

Theorem 5.5.5 (uniqueness of asymptotic expectation). Let $\{\xi_n\}$ be a sequence of random variables. Suppose that both $\frac{E\xi}{a_n}$ and $\frac{E\eta}{b_n}$ are asymptotic expectations of ξ_n . Then one of the following three must hold:

- (a) $E\xi = E\eta = 0$;
- (b) $E\xi \neq 0, E\eta = 0$, and $\frac{b_n}{a_n} \rightarrow 0$; or $E\xi = 0, E\eta \neq 0$, and $\frac{a_n}{b_n} \rightarrow 0$;
- (c) $E\xi \neq 0, E\eta \neq 0$, and $\frac{\frac{a_n}{E\xi}}{\frac{b_n}{E\eta}} \rightarrow 1$.

Proof. According to the definition, $a_n \xi_n \rightarrow_d \xi$ and $b_n \xi_n \rightarrow_d \eta$.

- (i) If both ξ and η has a nondegenerate c.d.f's, then the result
- (ii) Suppose that ξ has a nondegenerate c.d.f. but η is a constant. If $\eta \neq 0$, then by Slutsky's theorem $\frac{a_n}{b_n} \rightarrow \frac{\xi}{\eta}$, which is impossible since ξ has a nondegenerate c.d.f..
- (iii) Suppose that both ξ and η are constants. If $\xi = \eta = 0$, the result follows. If $\xi \neq 0$ and $\eta = 0$, then $\frac{b_n}{a_n} \rightarrow 0$. If $\xi \neq 0$ and $\eta \neq 0$, then $\frac{b_n}{a_n} \rightarrow \frac{\xi}{\eta}$. \square

5.5.3 Asymptotic Inference

Statistical inference based on asymptotic criteria and approximations is called asymptotic statistical inference. We now focus on asymptotic hypothesis tests and confidence sets.

Definition 5.5.6. Let $X = (X_1, \dots, X_n)$ be a sample from $P \in \mathcal{P}$ and $T_n(X)$ be a test for $H_0 : P \in \mathcal{P}_0$ versus $H_1 : P \in \mathcal{P}_1$.

- (i) If $\limsup_n \alpha_{T_n}(P) \leq \alpha$ for any $P \in \mathcal{P}_0$, then α is an asymptotic significance level of T_n .
- (ii) If $\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_0} \alpha_{T_n}(P)$ exists, then it is called the limiting size of T_n .

5.6 Unbisaed Estimation

5.7 Estimation in Parametric Models

5.8 Hypothesis tests

5.9 Confidence sets

Chapter 6

Multivariate Data Analysis

6.1 The Multivariate Normal Distributions

6.1.1 Asymptotic Distributions of Sample Means and Covariance Matrices

6.1.2 The Noncentral χ^2 and F Distributions

Definition 6.1.1 (generalized hypergeometric functions). The generalized hypergeometric function is

$${}_pF_q(a_1, \dots, a_p; b_1, \dots, b_q; z) = \sum_{n=0}^{\infty} \frac{(a_1)_n \cdots (a_p)_n}{(b_1)_n \cdots (b_q)_n} \frac{z^n}{n!}$$

where $(a)_k = a(a+1) \cdots (a+k-1)$

For our purpose we will make use of the results in the following two lemmas. The first gives a special integral for ${}_0F_1$.

Lemma 6.1.2.

$$\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n-1}{2})} \int_0^\pi e^{z \cos \theta} \sin^{n-2} \theta d\theta = {}_0F_1(\frac{n}{2}; \frac{z^2}{4})$$

The second lemma shows that ${}_{p+1}F_q$ is essentially a Laplace transform of ${}_pF_q$.

Lemma 6.1.3.

$$\int_0^\infty e^{-zt} t^{a-1} {}_pF_q(a_1, \dots, a_p; b_1, \dots, b_p; kt) dt = \Gamma(a) z^{-a} {}_{p+1}F_q(a_1, \dots, a_p, a; b_1, \dots, b_p; kz^{-1})$$

for $p < q$, $\Re(a) > 0$, $\Re(z) > 0$ or $p = q$, $\Re(a) > 0$, $\Re(z) > \Re(k)$.

Theorem 6.1.4. If X is $N_n(\mu, I_n)$ then the random variable $Z = X^T X$ has the density function

$$e^{-\frac{\delta}{2}} {}_0F_1(\frac{n}{2}; \frac{\delta z}{4}) \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} e^{-\frac{\delta}{2} z^{\frac{n}{2}-1}} \quad (z > 0)$$

where $\delta = \mu^T \mu$. Z is said to have the noncentral χ^2 distribution with n degrees of freedom and noncentrality parameter δ , to be written as $\chi_n^2(\delta)$.

Corollary 6.1.5. If Z is $\chi_n^2(\delta)$ then its density function can be expressed as

$$\sum_{k=0}^{\infty} P(K=k) g_{n+2k}(z) \quad (z > 0)$$

where K is a Poisson random variable with mean $\frac{\delta}{2}$, and

$$g_r(z) = \frac{1}{2^{\frac{r}{2}} \Gamma(\frac{r}{2})} e^{-\frac{\delta}{2} z^{\frac{r}{2}-1}}$$

the density function of the χ_r^2 distribution.

Theorem 6.1.6. If Z is $\chi_n^2(\delta)$ then its characteristic function is

$$\varphi_z(t) = (1 - 2it)^{\frac{n}{2}} e^{\frac{it\delta}{1-2it}}$$

Corollary 6.1.7. $E(Z) = n + \delta$ and $\text{Var}(Z) = 2n + 4\delta$

Corollary 6.1.8. If Z_1 is $\chi_{n_1}^2(\delta_1)$, Z_2 is $\chi_{n_2}^2(\delta_2)$, and Z_1 and Z_2 are independent, then $Z_1 + Z_2$ is $\chi_{n_1+n_2}^2(\delta_1 + \delta_2)$.

We now turn to the noncentral F distribution. Recall that the usual central F distribution is obtained by taking the ratio of two independent χ^2 variables divided by their degrees of freedom. The noncentral F distribution is obtained by allowing the numerator variable to be noncentral χ^2 .

Theorem 6.1.9. If Z_1 is $\chi_{n_1}^2(\delta)$, Z_2 is $\chi_{n_2}^2$, and Z_1 and Z_2 are independent, then

$$F = \frac{Z_1/n_1}{Z_2/n_2}$$

has the density function

$$e^{-\frac{\delta}{2}} {}_1F_1\left(\frac{n_1 + n_2}{2}; \frac{n_1}{2}; \frac{\frac{n_1}{2}\delta z}{1 + \frac{n_1}{n_2}z}\right) \times \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} \frac{z^{\frac{n_1}{2}-1} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}}{\left(1 + \frac{n_1}{n_2}z\right)^{\frac{n_1+n_2}{2}}} \quad z > 0$$

F is said to have the noncentral F distribution with n_1 and n_2 degrees of freedom and noncentrality parameter δ , to be written as $F_{n_1, n_2}(\delta)$.

Corollary 6.1.10. $E(F) = \frac{n_2(n_1+\delta)}{n_1(n_2-2)} (n_2 > 2)$ and $\text{Var}(F) = 2\left(\frac{n_2}{n_1}\right)^2 \frac{(n_1+\delta)^2 + (n_1+2\delta)(n_2-2)}{(n_2-2)^2(n_2-4)} (n_2 > 4)$

6.1.3 Quadratic Forms

Theorem 6.1.11. If X is $N_m(\mu, \Sigma)$, where Σ is nonsingular, then

- (i) $(X - \mu)^T \Sigma^{-1} (X - \mu)$ is χ_m^2
- (ii) $X^T \Sigma^{-1} X$ is $\chi_m^2(\delta)$ where $\delta = \mu^T \Sigma^{-1} \mu$

Theorem 6.1.12. If X is $N_m(\mu, I_m)$ and B is an $m \times m$ symmetric matrix, then $X^T B X$ has a noncentral χ^2 distribution if and only if B is idempotent.

6.1.4 Spherical and Elliptical Distributions

Definition 6.1.13. A $m \times 1$ random vector X is said to have a spherical distribution if X and OX have the same distribution for all $O \in O(n)$.

Theorem 6.1.14. Let X be $E_m(\mu, V)$, where V is diagonal. If X_1, \dots, X_m are all independent then X is normal.

Proof. WLOG assume $\mu = 0$. Then the characteristic function of X has the form $\phi(t) = \psi(t^T V t) = \psi(\sum_{i=1}^m t_i^2 v_{ii})$. Since X_1, \dots, X_m are independent we have This equation is known as Hamel's euqation and its only continuous solution is $\psi(z) = e^{kz}$ for some constant k . Hence the characteristic function of X has the form $\phi(t) = e^{kt^T V t}$, and because it is a characteristic function, we must have $k \leq 0$ which implies that X has a normal distribution. \square

6.2 Multivariate Integration

6.2.1 Exterior Products

For any matrix X , $d(X)$ denotes the matrix of differentials (dx_{ij}) .

For an arbitrary $n \times m$ matrix X , the symbol (dX) will denote the exterior product of the mn distinct elements of dX :

$$(dX) := \bigwedge_{j=1}^m \bigwedge_{i=1}^n dx_{ij}$$

For a symmetric $m \times m$ matrix X , the symbol (dX) will denote the exterior product of the $\frac{m(m+1)}{2}$ distinct elements of dX :

$$(dX) := \bigwedge_{1 \leq i \leq j \leq m} dx_{ij}$$

6.2.2 The Multivariate Gamma Function

6.2.3 Miscellaneous Jacobians

6.2.4 Invariant Measures

Chapter 7

von Mises Expansion

7.1 Int

Definition 7.1.1 (statistical functional). Let X_1, \dots, X_n be a sample from population with distribution function F and let $T_n = T_n(X_1, \dots, X_n)$ be a statistic. If T_n can be written as a functional T of the empirical distribution \hat{F}_n , i.e. $T_n = T(\hat{F}_n)$ where T does not depend on n .

Example 7.1.2. Let ϕ be a real-valued function and let

$$T_n = \frac{1}{n} \sum_{i=1}^n \phi(X_i).$$

Then for a general distribution G , the functional defined by

$$T(G) = \int \phi(x) dG(x)$$

satisfies $T_n(X_1, \dots, X_n) = T(\hat{F}_n)$. Functional of this form is called linear statistical functional.

Example 7.1.3. The sample k -th central moment $T_n = m_k = T_n(\hat{F}_n)$, where

$$T(G) = \int [x - \int x dG]^k dG(x)$$

7.2 Influence Function

Definition 7.2.1 (von Mises derivative). Let X_1, \dots, X_n be a sample and let T be a functional on a convex set of d.f.'s containing all empirical d.f.'s and the population d.f. F . Let G be a point in this convex set. Then the von Mises derivative T'_F of T at F is defined by

$$T'_F(G - F) = \frac{d}{dt} T(F + t(G - F))|_{t=0},$$

if there exists a real-valued function $\phi_F(x)$ (independent of G) s.t.

$$T'_F(G - F) = \int \phi_F(x) d(G - F)(x).$$

We shall normalize $\phi_F(x)$ by making

$$\int \phi_F(x) dF(x) = 0,$$

and call it the influence function.

7.3 Infinitesimal Jackknife

Jackknife is a general tool for estimating the variance and reducing the bias. The basic idea of Jackknife is to drop one observations and recompute the estimate using the remaining observations. This is done for each observation, and the results are then combined to produce an estimate.

If we assign each observation a weight, then dropping one observation is identical to giving it weight of zero. The Infinitesimal Jackknife

7.4 Perturbing a Training Input

How would the model's prediction change if a training input were modified?

Part I

Causal Inference

Chapter 8

Counterfactuals

This chapter aims to present the basic idea of counterfactuals in the language of probability theory. In this way, I hope to introduce new concepts and structures to enrich probability theory, and borrow tools and insights from probability theory to study causal inference. In the following I fix $(\Omega, \mathcal{F}, \mathbb{P})$ to be a probability space.

8.1 Binary Treatment

Let X, Y be random variables on Ω .

Example 8.1.1 (binary treatment). You can imagine Ω as the collection of all people being investigated, each element $\omega \in \Omega$ stands for a single person being investigated. Suppose X is a binary treatment variable, where $X = 1$ means 'treated' and $X = 0$ means 'not treated'. Let Y be some outcome variable such as the absence of disease. The goal is to study the relationship between Y and X .

Definition 8.1.2 (potential decomposition). If a random variable Y can be written as $Y = C_0 1_{A^c} + C_1 1_A$ where C_0 and C_1 are two random variables, then we say that Y admits a potential decomposition C_0, C_1 w.r.t. A .

Remark 8.1.3. If $X = 1_A$ is the binary treatment variable associated with A , then we also say that Y admits a potential decomposition w.r.t. X . In this case, we can call C_0, C_1 the potential outcomes with the following interpretation: C_0 is the outcome if not treated and C_1 is the outcome if treated.

Theorem 8.1.4 (existence). *For any random variable Y on Ω and any event $A \in \mathcal{F}$, there exists random variable C_0 and C_1 s.t.*

$$Y = C_0 1_{A^c} + C_1 1_A.$$

This theorem is self-evident. We can look at the following cases, where we assume $X = 1_A$ is a binary treatment variable.

Example 8.1.5. Let $C_0 = C_1 = Y$, then it is a potential decomposition of Y w.r.t. X . In this example, the outcome is the same whether treated or not. We can interpret this as X has no causal effect on Y .

Example 8.1.6. Let $C_0 = Y 1_D$ and $C_1 = Y 1_E$, where $A^c \subset D$, $A \subset E$ and $D, E \in \mathcal{F}$, then it is a potential decomposition of Y w.r.t. X . In this example, D, E can be chosen rather arbitrarily. This shows that potential decomposition is not unique.

The problem of the above example is that we can decompose a random variable *a posteriori*. To model the causal effect of the real world, we want the decomposition to be *a priori*. Namely, we are given a treatment X and potential outcomes C_0, C_1 first, then we construct Y naturally. We express this special type of potential decomposition more succinctly by

$$Y = C_X, \tag{8.1}$$

which is called the **consistency relationship**.

Now we can define statistics.

Definition 8.1.7 (average causal effect). Define the average causal effect or average treatment effect to be

$$\theta = \mathbb{E}C_1 - \mathbb{E}C_0. \tag{8.2}$$

8.2

Definition 8.2.1 (counterfactual function). A random variable which is parameterized by X .

Definition 8.2.2 (causal regression function). Define the causal regression function to be

$$\theta(x) = \mathbb{E}_\omega C(x, \omega). \quad (8.3)$$

Note that x is fixed.

Chapter 9

CPDAG

9.1 PDAGs

Definition 9.1.1 (PDAG). An acyclic partially directed graph, or PDAG for short, is a graph that contains both directed and undirected edges. It is acyclic in the sense that it contains no directed cycles.

Remark 9.1.2. An acyclic directed graph, or DAG for short, can be regarded as a PDAG without undirected edges.

Here we list some graph theoretical concepts for a PDAG.

- The *skeleton* of any PDAG is the undirected graph resulting from ignoring the directionality of every edge.
- A *v-structure* in a PDAG \mathcal{G} is an ordered triple of nodes (x, y, z) such that \mathcal{G} contains the directed edges $x \rightarrow y$ and $z \rightarrow y$, and x and z are not adjacent in \mathcal{G} .
- A *clique* in a PDAG is a set of nodes for which every pair of nodes is adjacent.

We also list some notations for a PDAG.

- We use Π_x to denote the set of parents for node x in a PDAG.
- We use N_x to denote the set of neighbors of node x in a PDAG.
- We use $N_{x,y} = N_x \cap N_y$ to denote the set of common neighbors of x and y in a PDAG.
- We use $\Omega_{x,y} = \Pi_x \cap N_y$ to denote the set of parents of node x that are neighbors of node y in a PDAG.

9.2 MECs

Two network structures are equivalent if the set of distributions that can be represented using one of the structures is identical to the set of distributions that can be represented using the other.

Theorem 9.2.1. *Two DAGs are Markov equivalent if and only if they have the same skeleton and the same v-structures.*

Proof.

□

Definition 9.2.2 (CPDAG). The CPDAG of a DAG \mathcal{D} , denoted as \mathcal{C} , is a PDAG that has the same skeleton as \mathcal{D} , and an edge is directed in \mathcal{C} if and only if it has the same orientation in every equivalent DAG of \mathcal{D} .

We can distinguish two kinds of edges in a DAG according to its CPDAG. A directed edge of a DAG is *compelled* if it occurs in the corresponding CPDAG, otherwise, the directed edge is *reversible*, and the corresponding parents are reversible parents.

9.3 A Transformational Characterization of MECs

Definition 9.3.1. A directed edge $x \rightarrow y$ is *covered* if $\Pi_y = \Pi_x \cup \{x\}$.

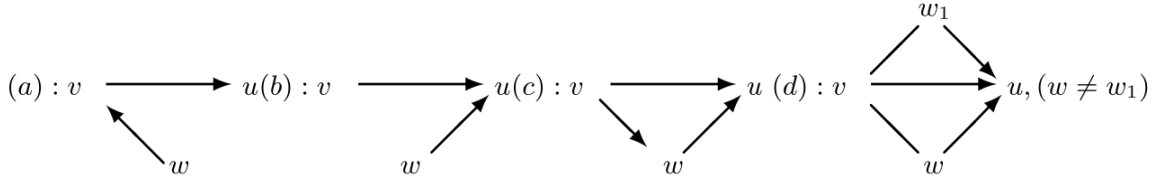
Theorem 9.3.2. Let \mathcal{G} be any DAG containing the edge $x \rightarrow y$, and let \mathcal{G}' be the directed graph identical to \mathcal{G} except that the edge between x and y in \mathcal{G}' is oriented as $y \rightarrow x$. Then \mathcal{G}' is a DAG that is equivalent to \mathcal{G} if and only if $x \rightarrow y$ is a covered edge in \mathcal{G} .

9.4 Characterization of CPDAGs

We introduce some terminologies and the main theorem.

- A PDAG is called a *chain graph* if it contains no semi-directed cycles.
- An undirected graph is *chordal* if every cycle of length greater than or equal to 4 possesses a chord.
- A directed edge $v \rightarrow u$ in a PDAG \mathcal{P} is *strongly protected* if $v \rightarrow u$ occurs in at least one of the four induced subgraphs of \mathcal{P} in Figure 9.1.

Figure 9.1: Configurations where $v \rightarrow u$ is strongly protected.



Theorem 9.4.1. A graph \mathcal{C} is a CPDAG of a directed acyclic graph \mathcal{D} if and only if \mathcal{C} satisfies the following properties:

1. \mathcal{C} is a chain graph;
2. let \mathcal{C}_τ be the subgraph induced by τ . \mathcal{C}_τ is chordal for every chain component τ ;
3. $w \rightarrow u - v$ does not occur as an induced subgraph of \mathcal{C} ;
4. every arrow $v \rightarrow u$ in \mathcal{C} is strongly protected.

Remark 9.4.2.

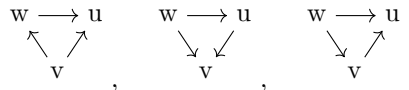
The following series of lemma closely inspect a clique of size three in a CPDAG \mathcal{C} , where there exists an undirected edge

Lemma 9.4.3. The configuration $w \rightarrow u - v$ does not occur as an induced subgraph of \mathcal{C} .

Proof. If $w \rightarrow u - v$ occurs as an induced subgraph in \mathcal{C} , then $w \rightarrow u \leftarrow v$ must occurs as a v-structure in some $\mathcal{D} \sim \text{Class}(\mathcal{C})$. By Theorem 9.2.1, this is impossible. \square

Lemma 9.4.4. If $\begin{array}{c} w \longrightarrow u \\ \diagdown \quad \diagup \\ v \end{array}$ occurs in \mathcal{C} , then there exist $\mathcal{D}_1, \mathcal{D}_2 \in \text{Class}(\mathcal{C})$ such that $\begin{array}{c} w \longrightarrow u \\ \diagdown \quad \diagup \\ v \end{array}$ occurs in \mathcal{D}_1 and $\begin{array}{c} w \longrightarrow u \\ \diagdown \quad \diagup \\ v \end{array}$ occurs in \mathcal{D}_2 .

Proof. Any $\mathcal{D} \in \text{Class}(\mathcal{C})$ must contain one of the followings:



\square

Lemma 9.4.5. *Let $\{w, u, v\}$ be any three nodes that form a clique of size three in \mathcal{C} . If any two of the edges in the clique are undirected, then the third edge is undirected as well.*

Proof. Assume $\begin{array}{c} w \longrightarrow u \\ \diagdown \quad \diagup \\ v \end{array}$ occurs in \mathcal{C} , we prove by contradiction. According to Lemma 9.4.4, suppose $\begin{array}{c} w \longrightarrow u \\ \diagdown \quad \diagup \\ v \end{array}$ occurs in \mathcal{D}_1 and $\begin{array}{c} w \longrightarrow u \\ \diagdown \quad \diagup \\ v \end{array}$ occurs in \mathcal{D}_2 .

As $w \rightarrow u$ is compelled, applying Theorem 9.3.2 to \mathcal{D}_1 yields that $w \rightarrow u$ is not a covered edge in \mathcal{D}_1 . Therefore, there exists x such that x is a parent of u but not a parent of w in \mathcal{G} . Now consider the subgraph induced by $\{u, v, w, x\}$ in \mathcal{D}_1 . If there is no edge between x and v , then $x \rightarrow u \leftarrow v$ forms a v-structure, which is impossible. So there must be an edge between x and v .

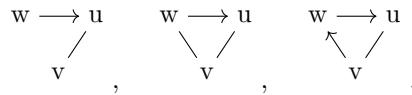
Now consider the subgraph induced by $\{u, v, w, x\}$ in \mathcal{D}_2 . As $x \rightarrow u \leftarrow w$ forms a v-structure in \mathcal{D}_1 , it also forms a v-structure in \mathcal{D}_2 by Theorem 9.2.1. Moreover, there must be an edge between x and v . If $v \rightarrow x$, then \mathcal{D}_2 contains a directed cycles. If $x \rightarrow v$ in \mathcal{D}_2 , then $x \rightarrow v \leftarrow w$ forms a v-structure, which is also impossible. □

Lemma 9.4.6. *Suppose there is a directed path from v to u in \mathcal{C} . If there is an edge between v and u , then it must be $v \rightarrow u$.*

Proof. Any $\mathcal{D} \in \text{Class}(\mathcal{C})$ must be acyclic. □

Corollary 9.4.7. *If $u - v$ is an undirected edge in \mathcal{C} , then $\Pi_u = \Pi_v$.*

Proof. Suppose not. Without loss of generality, let w be any parent of u that is not a parent of v . There are three configuration of the subgraph induced by $\{u, v, w\}$ in general:



The first case is excluded by Lemma 9.4.3, the second case is excluded by Lemma 9.4.5, and the third case is excluded by Lemma 9.4.6. □

Next we focus on the chain graph structure, *i.e.* maximal undirected components in \mathcal{C} . A maximal undirected component K of a CPDAG \mathcal{C} is a connected subgraph of \mathcal{C} , where each connecting edge is undirected, and for every node $x \in K$, if there is an undirected edge $y - x$ in \mathcal{C} , then $y \in K$. For the remainder of this note, we use *undirected component* to mean maximal undirected component.

Lemma 9.4.8. *For any directed edge $u \rightarrow v$ in \mathcal{C} , u is a parent of every node reachable by v via undirected edges.*

Proof. Follows by a repeated application of Corollary 9.4.7 along the edges in any undirected path. □

Lemma 9.4.9. *If there is a directed path of length one or more from u to v in \mathcal{C} , then u and v are not in the same undirected component.*

Proof. Suppose not. Consider the last edge $w \rightarrow v$ in the directed path. By Lemma 9.4.8, w is a parent of every node in the undirected component containing v , which means that w is a parent of u . This forms a cycle. □

The next lemma shows that the undirected components of a CPDAG are chordal.

Lemma 9.4.10. *\mathcal{C} has no undirected chordless k -cycles ($k \geq 4$).*

Proof. If an undirected chordless k -cycle occurs in \mathcal{C} , then \mathcal{C} must have at least one v-structure in this cycle. □


Corollary 9.4.11. *Let u and v be any pair of nodes that are not adjacent in \mathcal{C} . Then $N_{x,y}$ is a clique of undirected edges.*

Now we consider the converse of Corollary 9.4.7.

Lemma 9.4.12. *If $v \rightarrow u$ is a directed edge in \mathcal{C} , then $\Pi_u \neq \Pi_v \cup \{v\}$.*

Proof. □

Lemma 9.4.13. *Every directed edge $v \rightarrow u$ in \mathcal{C} is strongly protected.*

Proof. According to Lemma, we only need to show that if  occurs in \mathcal{C} , then the configuration (d) in Figure 9.1 occurs as an induced subgraph. □

Now we begin the proof of Theorem

Proof. We first prove the only if direction. □

9.5 Algorithm:

9.6 Algorithm:

Definition 9.6.1. An operator is valid if

- the modified graph of the operator is a PDAG and has a consistent extension;
-

9.6.1 Compelled and Reversible Edge Insertions

Theorem 9.6.2. *Let \mathcal{C} be any CPDAG for which nodes x and y are not adjacent. If the result of adding an edge between x and y is a PDAG that admits a consistent extension, then that edge is compelled if and only if $\Pi_x \neq \Pi_y$.*

One reason that Theorem 9.6.2 is difficult to prove is that after inserting edges into a CPDAG, edges that were reversible before can become compelled, and edges that were compelled before can become reversible. In the following lemmas, we demonstrate a method for determining which edges in a CPDAG will necessarily remain compelled and reversible after an edge addition, based on the topology of the CPDAG.

Lemma 9.6.3. *Let \mathcal{C} be a CPDAG, and let \mathcal{P} denote the PDAG that results from an edge addition or deletion. If \mathcal{P} admits a consistent extension, then any compelled edge $x \rightarrow y$ in \mathcal{C} cannot be compelled in the opposite direction in \mathcal{P} .*

Proof. □

9.6.2 The InsertU Operator

Theorem 9.6.4. *Let x and y be two nodes that are not adjacent in \mathcal{C} . The insertion of the undirected edge $x - y$ is valid if and only if (1) every undirected path between x and y contains a node in $N_{x,y}$, and (2) $\Pi_x = \Pi_y$.*

9.6.3 The DeleteU Operator

Theorem 9.6.5. *Let $x - y$ be an undirected edge in CPDAG \mathcal{C} . The deletion of $x - y$ is valid if and only if $N_{x,y}$ is a clique of undirected edges.*

9.6.4 The InsertD Operator

Lemma 9.6.6. *Let \mathcal{C} be a CPDAG that contains a semi-directed path from x to y . If there exists a directed edge $z \rightarrow w$ in this path, then there exists a directed path from z to y in \mathcal{C} .*

Theorem 9.6.7. *Let x and y be any two nodes that are not adjacent in CPDAG \mathcal{C} . The insertion of $x \rightarrow y$ is valid if and only if (1) every semi-directed path from y to x contains at least one node in $\Omega_{x,y}$, (2) $\Omega_{x,y}$ is a clique of undirected edges, and (3) $\Pi_x \neq \Pi_y$.*

9.6.5 The DeleteD Operator

Theorem 9.6.8. *Let $x \rightarrow y$ be a directed edge in CPDAG \mathcal{C} . The deletion of $x \rightarrow y$ is valid if and only if N_y is a clique of undirected edges.*

9.6.6 The ReverseD Operator

Theorem 9.6.9. *Let $x \rightarrow y$ be a directed edge in CPDAG \mathcal{C} . The reversal of $x \rightarrow y$ is valid if and only if (1) every semi-directed path from x to y that does not include the edge $x \rightarrow y$ contains at least one node in $\Omega_{y,x} \cup N_y$, and (2) $\Omega_{y,x}$ is a clique of undirected edges.*

9.6.7 The MakeV Operator

Theorem 9.6.10. *Let $x - z - y$ be any length-two undirected path in CPDAG \mathcal{C} such that x and y are not adjacent. Replacing the undirected edges with directed edges to create the v-structure $x \rightarrow z \leftarrow y$ is valid if and only if every undirected path between x and y contains a node in $N_{x,y}$.*

9.7 Algorithm: Reversible MCMC on MECs

We need to show each operator has a reversible operator.

9.7.1 The InsertU Operator

Theorem 9.7.1. *Let x and y be two nodes that are not adjacent in \mathcal{C} . The insertion of the undirected edge $x - y$ is reversible if and only if (1) every undirected path between x and y contains a node in $N_{x,y}$, and (2) $\Pi_x = \Pi_y$.*

9.7.2 The DeleteU Operator

Theorem 9.7.2. *Let $x - y$ be an undirected edge in CPDAG \mathcal{C} . The deletion of $x - y$ is valid if and only if $N_{x,y}$ is a clique of undirected edges.*

9.7.3 The InsertD Operator

Lemma 9.7.3. *Let \mathcal{C} be a CPDAG that contains a semi-directed path from x to y . If there exists a directed edge $z \rightarrow w$ in this path, then there exists a directed path from z to y in \mathcal{C} .*

Theorem 9.7.4. *Let x and y be any two nodes that are not adjacent in CPDAG \mathcal{C} . The insertion of $x \rightarrow y$ is valid if and only if (1) every semi-directed path from y to x contains at least one node in $\Omega_{x,y}$, (2) $\Omega_{x,y}$ is a clique of undirected edges, and (3) $\Pi_x \neq \Pi_y$.*

9.7.4 The DeleteD Operator

Theorem 9.7.5. *Let $x \rightarrow y$ be a directed edge in CPDAG \mathcal{C} . The deletion of $x \rightarrow y$ is valid if and only if N_y is a clique of undirected edges.*

9.7.5 The RemoveV Operator

First we show that RemoveC is valid.

Theorem 9.7.6.

Next we show that RemoveC is reversible.

9.8 Counting Sizes of MECs

9.8.1 Size of MEC Determined by the Number of Vertices

Let $\mathcal{U}_{p,n}$ be an undirected and connected chordal graph (UCCG) with p vertices and n edges.