

Probability and Statistics

PhD Qualification Exam

Mock Examination No. 5 – Solutions

Problem 1

10 points

两个向量都是 X 的线性变换，故 jointly Gaussian。其交叉协方差为

$$\text{Cov}(PX, (I - P)X) = P(I - P) = 0.$$

Jointly Gaussian vectors 零交叉协方差即独立。

取正交矩阵将 P 对角化为 $\text{diag}(I_r, 0)$ 。正交不变性给出

$$\|PX\|^2 \sim \chi_r^2, \quad \|(I - P)X\|^2 \sim \chi_{n-r}^2,$$

且二者独立。因此

$$\frac{\|PX\|^2/r}{\|(I - P)X\|^2/(n - r)} \sim F_{r, n-r}.$$

独立 Gamma variables 的比例性质还给出

$$\frac{\|PX\|^2}{\|X\|^2} \sim \text{Beta}\left(\frac{r}{2}, \frac{n - r}{2}\right).$$

Scoring guide (10 points). 联合高斯和交叉协方差 2 分、独立性 1 分；正交对角化 1 分、两个独立卡方 2 分；F 分布 2 分、Beta 分布 2 分。

Problem 2

15 points

令 $Y_{nk} = a_{nk}X_k$ 。每行变量独立、中心化，且总方差为 $\sum_k a_{nk}^2 = 1$ 。Lyapunov quantity 满足

$$\sum_{k=1}^{m_n} \mathbb{E}|Y_{nk}|^{2+\delta} = \mathbb{E}|X_1|^{2+\delta} \sum_k |a_{nk}|^{2+\delta} \leq \mathbb{E}|X_1|^{2+\delta} \left(\max_k |a_{nk}|\right)^\delta \sum_k a_{nk}^2 \rightarrow 0.$$

所以 Lyapunov triangular-array CLT 的全部条件满足，得到

$$\sum_{k=1}^{m_n} a_{nk}X_k \xrightarrow{d} N(0, 1).$$

若删除最大权重条件，取 $m_n = 1, a_{n1} = 1$ ，并取任一中心化、方差一、有限所有矩但非正态的 X_1 ，例如 Rademacher variable。则每个和恒等于 X_1 ，不可能趋于标准正态。

更弱的常用替代是 Lindeberg condition：对每个 $\varepsilon > 0$,

$$\sum_k \mathbb{E}[Y_{nk}^2 \mathbf{1}_{\{|Y_{nk}| > \varepsilon\}}] \rightarrow 0.$$

在总方差趋于一旦 infinitesimal 条件下，Lindeberg–Feller CLT 即可使用，无需 $(2 + \delta)$ moment。

Scoring guide (15 points). 三角阵列设定和总方差 2 分、Lyapunov 上界 4 分、定理与结论 3 分；合法反例 4 分；Lindeberg condition 2 分。

Problem 3**10 points**

Tower property 给出

$$\mathbb{E}(M_{n+1} \mid \mathcal{F}_n) = \mathbb{E}\{\mathbb{E}(H \mid \mathcal{F}_{n+1}) \mid \mathcal{F}_n\} = M_n,$$

故为 martingale。Conditional Jensen inequality 给出 $\mathbb{E}M_n^2 \leq \mathbb{E}H^2$, 所以 L^2 bounded。若 $j > i$, 则

$$\mathbb{E}[(M_{i+1} - M_i)(M_{j+1} - M_j)] = 0,$$

因为较早增量是 $\mathcal{F}_j$ -measurable, 而较晚增量给定 $\mathcal{F}_j$ 的条件均值为零。于是 increments 在 L^2 正交, Pythagorean identity 给出

$$\mathbb{E}M_n^2 = \mathbb{E}M_1^2 + \sum_{k=1}^{n-1} \mathbb{E}(M_{k+1} - M_k)^2.$$

在 Hilbert space L^2 中, M_n 是 H 到闭子空间 $L^2(\mathcal{F}_n)$ 的正交投影。这些子空间递增, 其并的闭包是 $L^2(\mathcal{F}_\infty)$ 。Nested projection theorem 因而给出

$$M_n \longrightarrow \text{Proj}_{L^2(\mathcal{F}_\infty)} H = \mathbb{E}(H \mid \mathcal{F}_\infty) \quad \text{in } L^2.$$

Scoring guide (10 points). tower property/martingale 2 分、Jensen 有界 1 分; 增量正交 2 分、平方和恒等式 1 分; 投影视角 2 分、极限空间识别和 L^2 结论 2 分。

Problem 4**15 points**由 Fubini theorem 与 Brownian covariance $\text{Cov}(B_s, B_t) = \min(s, t)$,

$$\text{Var}(A) = \int_0^1 \int_0^1 \min(s, u) \, ds \, du = \frac{1}{3},$$

以及

$$\text{Cov}(B_t, A) = \int_0^1 \min(t, s) \, ds = t - \frac{t^2}{2} =: c(t).$$

 (B_t, A) jointly Gaussian, 所以 Gaussian conditioning formula 给出

$$B_t \mid A = a \sim N(3c(t)a, t - 3c(t)^2).$$

定义

$$W_t = B_t - 3c(t)A.$$

则 $\text{Cov}(W_t, A) = c(t) - 3c(t)\text{Var}(A) = 0$ 。整个 $(W_t)_{0 \leq t \leq 1}$ 与 A jointly Gaussian, 故所有有限维向量与 A 独立。其协方差为

$$\text{Cov}(W_s, W_t) = \min(s, t) - 3c(s)c(t).$$

且 $B_t = W_t + \mathbb{E}(B_t \mid A)$ 。

取 $t = 1$, $c(1) = 1/2$, 故

$$B_1 | A = a \sim N\left(\frac{3a}{2}, \frac{1}{4}\right), \quad \mathbb{P}(B_1 > 0 | A = a) = \Phi(3a).$$

Scoring guide (15 points). 双重积分/方差 2 分、协方差 2 分；条件均值和方差各 2 分；残差过程 1 分、独立性 1 分、协方差 2 分； $t = 1$ 条件律 1 分、概率 2 分。

Problem 5

15 points

平移后积分为

$$\int_{\theta}^{\theta+1} 2(x - \theta) dx = 1.$$

样本 likelihood 是

$$L(\theta; x) = 2^n \prod_{i=1}^n (x_i - \theta) \mathbf{1}_{\{x_{(n)} - 1 < \theta < x_{(1)}\}}.$$

由 dominated-family minimal-sufficiency likelihood-ratio criterion, 两样本 x, y 属于同一 statistic fiber 当且仅当 $L(\theta; x)/L(\theta; y)$ 与 θ 无关。若两个 support intervals

$$(x_{(n)} - 1, x_{(1)}) \quad \text{和} \quad (y_{(n)} - 1, y_{(1)})$$

不同, 则 ratio 在某个参数区间从正值变为零或反之, 不可能为常数。若区间相同, 两个端点相同; 在共同区间内 ratio 是两个首一多项式 $\prod_i (x_i - \theta)$ 与 $\prod_i (y_i - \theta)$ 之比。它若为常数, 比较最高次项知常数为 1, 再由 polynomial identity 得两者根的 multiset 相同, 即两组 order statistics 完全相同。反向结论显然成立。因此 T minimal sufficient。

令 $R = X_{(n)} - X_{(1)}$ 。因为 $X_i - \theta$ 的共同分布不依赖 θ , R 是 ancillary; 同时 R 是 T 的有界、非退化函数。令 $c = \mathbb{E}_{\theta} R$ (它与 θ 无关), 则

$$g(T) = R - c$$

对所有 θ 均值为零但不 a.s. 为零。因此 T 不 complete。

Scoring guide (15 points). 归一化 1 分、完整 likelihood/support 2 分; ratio criterion 2 分、support interval 必要性 2 分、多项式恒等式 3 分、充要结论 1 分; ancillary range 2 分、零均值非零函数与不完全性 2 分。

Problem 6

10 points

令 $K = \sum_i \mathbf{1}_{\{X_i \leq m\}}$ 。连续性与 $F(m) = 1/2$ 给出 $K \sim \text{Binomial}(n, 1/2)$ 。采用题面的无界端点约定。对整数 $0 \leq k \leq \lfloor n/2 \rfloor$,

$$\{X_{(k)} \leq m \leq X_{(n-k+1)}\} = \{k \leq K \leq n - k\}$$

除零概率边界外成立。因此选择最大的 k 使

$$\sum_{j=k}^{n-k} \binom{n}{j} 2^{-n} \geq 1 - \alpha$$

($k = 0$ 总满足, 并给出平凡区间 $(-\infty, \infty)$), 则 $[X_{(k)}, X_{(n-k+1)}]$ 是 distribution-free coverage 至少 $1 - \alpha$ 的区间。需要 exact equality 时可在相邻 k 规则之间随机化。

若 f 在 m 连续且 $f(m) > 0$, sample quantile CLT (可由 empirical CDF CLT 和局部线性展开得到) 给出

$$\sqrt{n}(\hat{m} - m) \xrightarrow{d} N\left(0, \frac{1}{4f(m)^2}\right).$$

Scoring guide (10 points). Binomial count 2 分、order-statistic event 2 分、index/coverage 2 分; quantile CLT 条件 2 分、方差与结论 2 分。

Problem 7

15 points

Multinomial likelihood 在 simplex 上由

$$\hat{p}_i = \frac{N_i}{n}, \quad i = 1, 2, 3$$

最大化。故

$$-2 \log \Lambda = 2 \sum_{i=1}^3 N_i \log \frac{N_i}{np_{i0}} =: G^2,$$

并约定 $0 \log 0 = 0$ 。

零假设参数是二维 simplex 的 interior point (所有 $p_{i0} > 0$), multinomial model 光滑且 information 在二维切空间上非奇异。Unrestricted model 与 point null 的维数差为 2, 故 Wilks theorem 给出

$$G^2 \xrightarrow{d} \chi_2^2.$$

令 $\Delta_i = N_i - np_{i0}$ 。在 H_0 下 $\Delta_i = O_p(\sqrt{n})$, 且

$$\log \left(1 + \frac{\Delta_i}{np_{i0}}\right) = \frac{\Delta_i}{np_{i0}} - \frac{\Delta_i^2}{2n^2 p_{i0}^2} + O_p(n^{-3/2}).$$

代入 G^2 , 使用 $\sum_i \Delta_i = 0$, 线性项消失, 二次项给出

$$G^2 = \sum_{i=1}^3 \frac{\Delta_i^2}{np_{i0}} + o_p(1) = X^2 + o_p(1).$$

因此两个统计量均可用规则 “大于 $\chi_{2,1-\alpha}^2$ 时拒绝”, 得到 asymptotic level α 。

Scoring guide (15 points). MLE 2 分、LR 推导 2 分、零计数约定 1 分; Wilks 维数 2 分、正则条件 1 分、极限 1 分; Taylor 展开 2 分、线性抵消 2 分、 $o_p(1)$ 与规则 2 分。

Problem 8**10 points**

对任意实数 z 和任意 $\theta \in [a, b]$, 到闭凸集 $[a, b]$ 的投影满足

$$|\Pi_{[a,b]}(z) - \theta| \leq |z - \theta|.$$

令 $z = \delta(X)$, 平方后逐点成立, 再取 $\mathbb{E}_\theta$ 得

$$R(\theta, \delta^*) \leq R(\theta, \delta) \quad \text{对每个 } \theta \in [a, b].$$

当 $\mathbb{P}_\theta\{\delta \notin [a, b]\} > 0$ 时, 上述距离不等式在该事件上严格, 因而只要风险有限, 就有 $R(\theta, \delta^*) < R(\theta, \delta)$ 。如果对至少一个参数严格、对所有参数不增, 则 δ^* 严格支配 δ 。所以一个 admissible estimator 不应在具有正概率的样本上取到参数区间之外; 否则其投影会支配它 (忽略所有参数下均为零的样本集)。

Scoring guide (10 points). 逐点投影不等式 3 分、平方取期望得到风险比较 2 分; 严格性条件 2 分、strict domination 1 分、admissibility 后果 2 分。