Computational Statistics

Max Chen

Abstract

This is the summary notes of *Computational Statistics*, based on the lecture notes of Prof Geoff Nicholls and Prof Frank Windmeijer. Most of the contents are directly from the the lecture notes and slides of *Computational Statistics*, although some reordering and rearrangements are made in sake of helping the readers to understand the materials. Some gaps (e.g. proofs and derivations) are filled to the original materials, based on the handwritten annotations during the lectures. Some personal ideas are also added, hence may not be 100% theoretically rigorous but should be helpful for the comprehension of the materials. Some Proofs and examples are omitted, hence again the readers are recommended to read the original notes for details.
No person or party should use this notes for any purpose other than studying and understanding the notes itself.

Contents

1 Non-Parametric Rank Tests

1.1 Permutation Test

1.1.1 Exchangeability

Exchangeable: $X_1, \cdots, X_n \sim X_{\sigma_1}, \cdots, X_{\sigma_n}$, for any permutation $\sigma \in \mathcal{P}_n$.
- Exchangeability $\implies p_{1:n}(x_1, \cdots, x_n) \overset{d}{=} p_{1:n}(x_{\sigma_1}, \cdots, x_{\sigma_n}), \forall \sigma \in \mathcal{P}_n$.

Prop 1.2: $X_{1:n}$ exchangeable $\implies Pr(X \leq a) = Pr(X \leq a_\sigma)$.
Proof:

$$\begin{aligned} RHS = \operatorname{Pr}(X \leq a_\sigma) &= \int_{x_i \leq a_{\sigma_i}, i=1,\ldots,n} p_{1:n}(x_1, \ldots, x_n)\, dx_1 \ldots dx_n \\ &= \int_{\tilde{x}_{\sigma_i} \leq a_{\sigma_i}, i=1,\ldots,n} p_{1:n}(\tilde{x}_{\sigma_1}, \ldots, \tilde{x}_{\sigma_n})\, d\tilde{x}_{\sigma_1} \ldots d\tilde{x}_{\sigma_n}, \text{ by CoV } x_i \to \tilde{x}_{\sigma_i} \text{ with Jacobian} = 1 \\ &= \int_{\tilde{x}_i \leq a_i, i=1,\ldots,n} p_{1:n}(\tilde{x}_1, \ldots, \tilde{x}_n)\, d\tilde{x}_1 \ldots d\tilde{x}_n, \text{ by exchangeability}(\bullet) \\ &= \operatorname{Pr}(X \leq a) = LHS \quad \text{[EOP]} \end{aligned}$$

1.1.2 Ranks

Let $Z = (X, Y)$ with $Z_{1:m} = X_{1:m}$ and $Z_{m+1:N} = Y_{1:n}$.
Rank of Z_i: suppose without **tie** ($Z_i \neq Z_j, \forall i, j$),

$$R_i = R(Z_i) = \sum_{j=1}^{N} \mathbb{I}\{Z_j \leq Z_i\}$$

with the **rank function** $R(Z) = (R(Z_1), \cdots, R(Z_N))$.
- The rank function defines the order statistics $Z_{(R)} = (Z_{R_1}, \cdots, Z_{R_N})$.

Prop 1.4: If $Z_{1:N}$ is exchangeable, then $R(Z) \sim U(\mathcal{P}_N)$ with $P(R(Z) = r) = \frac{1}{N!}$.
Proof: Let $d_i = (j : r_j = i)$ be the set of the original index of the ordered statistic. ($d = r^{-1}$ mapping from order back to index).

$$\begin{aligned} P(R(Z) = r) &= P(Z_{d_1} < Z_{d_2} < \ldots < Z_{d_N}) \\ &= \int_{z_{d_1} < z_{d_2} < \ldots < z_{d_N}} p_{1:N}(z_1, \ldots, z_N)\, dz_1 \ldots dz_N \\ &= \int_{\tilde{z}_1 < \tilde{z}_2 < \ldots < \tilde{z}_N} p_{1:N}(\tilde{z}_{r_1}, \ldots, \tilde{z}_{r_N})\, d\tilde{z}_{r_1} \ldots d\tilde{z}_{r_N}, \overset{CoV}{\Longleftarrow} \tilde{z}_i \leftarrow z_{d_i} \text{ s.t. } z_j = \tilde{z}_{r_j} \text{ with Jacobian=0} \\ &= \int_{\tilde{z}_1 < \tilde{z}_2 < \ldots < \tilde{z}_N} p_{1:N}(\tilde{z}_1, \ldots, \tilde{z}_N)\, d\tilde{z}_1 \ldots d\tilde{z}_N, \text{ by exchangeability} \\ &= P(R(Z) = (1, 2, \ldots, N)), \text{ hence equal probability for any r, [EOP]} \end{aligned}$$

- $r : \mathcal{P}_N \to \mathcal{P}_N$ (from index to rank) and $d : \mathcal{P}_N \to \mathcal{P}_N$ (from rank to index).

Prop 1.5: $Z_{1:N}$ exchangeable $\implies q\left(z_{(1)}, \ldots, z_{(N)}\right) = N! p(z_1, \ldots, z_N)$, where $q(z_{(1)}, \cdots, z_{(N)})$ is the joint density of the order statistic.

Prop 1.6: $Z_{1:N}$ exchangeable $\implies R(Z) \perp (Z_{(1)}, \cdots, Z_{(N)})$.
<u>Proof</u>: Note the 2 relations,

$$
\begin{aligned}
&\textcircled{1} p\left(z_1, \ldots, z_N\right)=q\left(z_{(1)}, \ldots, z_{(N)}\right) P(R(Z)=r) \\
&\textcircled{2} p\left(z_1, \ldots, z_N\right)=p\left(z_1, \ldots, z_N \mid Z_{d_1}<Z_{d_2}<\ldots<Z_{d_N}\right) P\left(Z_{d_1}<Z_{d_2}<\ldots<Z_{d_N}\right) \\
&\qquad =p\left(z_1, \ldots, z_N \mid R(Z)=r\right) P(R(Z)=r)=p\left(z_{(r_1)}, \ldots, z_{(r_N)} \mid R(Z)=r\right) P(R(Z)=r) \\
&\qquad \Longrightarrow p\left(z_{(r_1)}, \ldots, z_{(r_N)} \mid R(Z)=r\right)=q\left(z_{(1)}, \ldots, z_{(N)}\right)=p\left(z_{(r_1)}, \ldots, z_{(r_N)}\right) \quad \text{[EOP]}
\end{aligned}
$$

1.1.3 Distribution free test statistics

Let $\mathcal{D}$ be a set of probability distributions and $Z = (Z_1, \cdots, Z_N) \sim D \in \mathcal{D}$.
A test statistics $T(Z)$ is **distribution free** over $\mathcal{D}$ if $T(Z)$ has the same distribution $\forall D \in \mathcal{D}$.
• If $T(Z)$ is a function of $R(Z)$, then exchangeability of $Z \implies$ distribution free $T(Z)$.

1.1.4 Two-Sample Permutation Test

2-Sample Testing: 2 samples $(X_i)_{i=1}^m \overset{iid}{\sim} F_X$ and $(Y_j)_{j=1}^n \overset{iid}{\sim} F_Y$ jointly independent, with F_X, F_Y arbitrary,

$$H_0 : F_X = F_Y \quad VS \quad H_1 : F_X \neq F_Y$$

2-Sample Permutation Testing: a 2-Sample testing with,
(i) *test statistic*: $T(Z) \mid Z_{(1)} = z_{(1)}, \ldots, Z_{(N)} = z_{(N)}$, with $T(Z) = T(z_{(R_Z)})$.
(ii) *p-value*: (reject H_0 when T_{obs} is large)

$$
\begin{aligned}
p &= \Pr\left(T(Z) \geq T_{obs} \mid Z_{(1)} = z_{(1)}, \ldots, Z_{(N)} = z_{(N)}, H0\right) \\
&= E\left(\mathbb{I}_{T\left(z_{(R)}\right) \geq T_{obs}}\right), \text{by Prop 1.7} \\
&= \frac{1}{N!} \sum_{r \in \mathcal{P}_N} \mathbb{I}\left\{T\left(z_r\right) \geq T_{obs}\right\}, \quad \because R \sim U\left(\mathcal{P}_N\right)
\end{aligned}
$$

(iii) *assumption*: joint independence of all entries in $Z = (X, Y)$. (rejecting H_0 means assumption fails, as X, Y come from the same distribution.)

Prop 1.7: If $Z_{1:N}$ exchangeable and $R \sim U(\mathcal{P}_N)$ (uniform random permutation), then

$$T(Z) | Z_{(1)}, \cdots, Z_{(N)} \overset{d}{=} T(z_{(R)})$$

with $z_{(R)} = \left(z_{(R_1)}, \cdots, z_{(R_N)}\right)$
<u>Proof</u>: Define $T(Z) := T'\left(Z_{(1)}, \ldots, Z_{(N)}, R(Z)\right)$ $(\because Z = Z_{(R(Z))})$

$$
\begin{aligned}
T(Z) \mid Z_{(1)} = z_{(1)}, \ldots, Z_{(N)} = z_{(N)} &= T'\left(z_{(1)}, \ldots, z_{(N)}, R(Z)\right) \mid Z_{(1)} = z_{(1)}, \ldots, Z_{(N)} = z_{(N)} \\
&= T'\left(z_{(1)}, \ldots, z_{(N)}, R(Z)\right), \quad \because \text{ order statistic is given} \\
&= T(z_{(R_Z)}), \quad \because R(Z) \perp Z_{(1):(N)} \text{ and } T(z_{(R_Z)}) \sim T(Z_{(R)}) \quad \text{[EOP]}
\end{aligned}
$$

Choice of Test Statistics: need some prior knowledge of how F_X and F_Y may differ, e.g. the different in means,

$$T(Z) = \left| \frac{1}{n} \sum_{i=1}^{n} Z_i - \frac{1}{m} \sum_{j=n+1}^{n+m} Z_j \right| = |\bar{X} - \bar{Y}|$$

• This is great when 2 samples are different in mean but relatively similar variance, but this can be a terrible statistic when 2 samples are different in variance but similar in mean.
(see Example 1.8 at P8 on the notes.)

1.1.5 Monte Carlo Permutation Test

• When N is large, we are unable to compute the p-value as it involves $N!$.

```
Algorithm 1: Monte Carlo Permutation Test
  Initialize the number of ranks to simulate $K$.
  Initialize the test statistic $T(Z)$ and significance level $\alpha$.
  for $k = 1, \cdots, K$ do
      Simulate $r^{(k)} \sim U(\mathcal{P}_N)$ with $r^{(k)} = \left(r_1^{(k)}, \cdots, r_N^{(k)}\right)$;
      Compute test statistic: $T^{(k)} = T(z_{(r^{(k)})})$
  end
  Compute the approximated p-value: $\hat{p} = K^{-1} \sum_{k=1}^{K} \mathbb{I}\left\{T^{(k)} \geq T_{obs}\right\}$
  Reject $H_0$ if $\hat{p} \leq \alpha$
```

1.1.6 Invariance Under Monotone Transformations

Monotone transformations don't affect the ranks but they do affect the order-statistics. So, as long as the permutation test statistic is conditioned on the order statistics, its distribution changes wrt monotone transformations.
(see Example 1.10 at P10-11 on the notes.)

1.2 Wilcoxon Rank Sum Test

Wilcoxon Rank Sum Test A 2-sample test for *location*, with,
(i) *Test assumption*: $(X_i)_{i=1}^m \sim F$ and $(Y_j - \Delta)_{j=1}^n \sim F$ jointly independent,
- with Δ the **location shift** between X and Y;
(ii) *Hypothesis*: $H_0 : \Delta = 0$ VS $H_1 : \Delta \neq 0$ or $\Delta > 0$ or $\Delta < 0$;
• Under H_0, $\Delta = 0$ and $X, Y \overset{iid}{\sim} F$.
(iii) *Test Statistics*: $W(R(Z)) = \sum_{i=m+1}^{N} R_i$, with $N = n + m$;
(iv) *p-values*:
- **positive shift**: $H_1 > 0$ compute $p^+ = \Pr\left(W(R) \geq W_{obs}\right)$ and reject H_0 if $p^+ \leq \alpha$;
- **negative shift**: $H_1 < 0$ compute $p^- = \Pr\left(W(R) \leq W_{obs}\right)$ and reject H_0 if $p^- \leq \alpha$;
- **non-zero shift**: $H_1 \neq 0$ compute p^+, p^- and reject H_0 if $2\min(p^+, p^-) \leq \alpha$ (by Prop 1.8).

Prop 1.8: The distribution of W(R) is symmetric about $\mu_W = \frac{n(m+n+1)}{2}$, and has variance $V_W = \frac{mn(m+n+1)}{2}$. (see PS1 or the proof below)
Proof: Denote $\mathcal{W}_{m,n}$ be the sample space of $W(R)$, then $\mathcal{W}_{m,n} = \{a, a+1, \cdots, b\}$, where:
- $a = \frac{n(n+1)}{2}$, when the last n are the smallest;
- $b = \frac{(m+n)(m+n+1)}{2} - \frac{m(m+1)}{2}$, when the last n are the largest.
Note that $W \sim U(\mathcal{W}_{m,n})$, as $R \sim U(\mathcal{P}_N)$.
Hence $\mu_W = E(W(R)) = \frac{a+b}{2} = \frac{n(m+n+1)}{2}$ and $V_W = \frac{n^2-1}{2} = \frac{(b-a+1)^2-1}{2} = \frac{mn(m+n+1)}{2}$ [EOP].

1.2.1 Mann-Witney test statistic

An equivalent test statistic of the Wilcoxon Rank Sum test statistic,

$$U(X, Y) = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{I}\left\{Y_j > X_i\right\} = W(R(Z)) - \frac{n(n+1)}{2}$$

Proof: (omitted, see PS1)

• Wilcoxon is differed from Mann-Witney by assuming location shift (and all else equal in the distribution of X, Y) under H_0, whereas Mann-Witney assumes $F_X \neq F_Y$ completely.

1.2.2 Test by Quantile

Suppose $G_{m,n}(w_\alpha) = \alpha$, with significant level $\alpha > 0$, then reject H_0:
- **lower tail** if $W_{obs} \leq w_\alpha$ under alternative $H1 : \Delta < 0$;
- **upper tail** if $W_{obs} \geq w_{1-\alpha}$ under alternative $H1 : \Delta > 0$;
- **equal tail** if $W_{obs} \leq w_{\alpha/2}$ or $W \geq w_{1-\alpha/2}$ under $H1 : \Delta \neq 0$;

where $G_{m,n}(w) = Pr(W(R) \leq w)$ is defined as the CDF of Wilcoxon Rank Sum Test
• Because by Prop 1.8 (symmetry of $W(R)$), if $a < w \leq \mu_W$ and $\delta = \mu_W - w$, then $\Pr\left(W(R) \leq \mu_W - \delta\right) = \Pr\left(W(R) \geq \mu_W + \delta\right)$.

1.2.3 The Monte Carlo Wilcoxon Rank Sum Permutation Test

• For large m, n, going through the Wilcoxon Rank Sum Test becomes impossible, hence looking for an approximation of $G_{m,n}$.

```
Algorithm 2: Monte Carlo Wilcoxon Rank Sum Permutation Test
  Initialize the test statistic $W(R)$ as above.
  Initialize the number of ranks $K$ to simulate and the significance level $\alpha$.
  for $k = 1, \cdots, K$ do
      Simulate $R^{(k)} \overset{iid}{\sim} U(\mathcal{P}_{m+n})$;
      Compute $W^{(k)} = W(R^{(k)})$.
  end
  Compute p-value: $\hat{p} = \begin{cases} \hat{p}^{+} = K^{-1} \sum_k \mathbb{I}_{W^{(k)} \geq W_{obs}}, & \text{for upper tail} \\ \hat{p}^{-} = K^{-1} \sum_k \mathbb{I}_{W^{(k)} \leq W_{obs}}, & \text{for lower tail} \\ 2\min(\hat{p}^{+}, \hat{p}^{-}), & \text{for equal tail} \end{cases}$
  Reject $H_0$ if $\hat{p} \leq \alpha$
```

1.2.4 Problem with Ties

Everything above fails if we have ties in the data, because Z can have different ranks $R(Z)$ (hence different $W(R(Z))$) depending on the order the data is observed.
However, the Monte Carlo permutation test works if we use the same tie-rule to compute the simulated W. We can deal with tied-data by: permute the data Z, compute rank $R(Z)$ and test statistic $W(R(Z))$, then test $W \leq W_{obs}$ or $W \geq W_{obs}$.

1.2.5 Normal Approximation to the Distribution of W

By CLT,

$$\frac{W - \mu_W}{\sqrt{V_W}} \xrightarrow{d} \mathcal{N}(0,1) \quad n, m \to \infty$$

where μ_W, V_W from Prop 1.8.

Continuity correction: since test statistic is discrete (as ranks taking integers) and normal distribution is continuous, we do Continuity correction by:
- take a small step to the right (e.g. $+\frac{1}{2}$) when estimating $\hat{p}^-$ for the lower tail; OR
- take a small step to the left (e.g. $+\frac{1}{2}$) when estimating $\hat{p}^+$ for the upper tail; OR
• Both lead to larger p-values, hence more conservative and less likely to reject H_0.
(see example 1.14-1.15 at P18 on the notes.)

1.2.6 Estimation of Location Δ

2-Sided z-Test to estimate location:
(i) *Assumption*: $X \sim N(0, \sigma^2), Y \sim N(\Delta, \sigma^2)$;
(ii) *Hypothesis*: $H_0 : \Delta = \tilde{\Delta}$ VS $H_1 : \Delta \neq \tilde{\Delta}$;
(iii) *Test statistic*: $Z(\tilde{\Delta}) = \frac{(\bar{Y} - \tilde{\Delta} - \bar{X})}{\sigma\sqrt{1/n + 1/m}} \overset{CLT}{\sim} N(0,1)$;
(iv) *p-value*: $2(1 - \Phi(|Z(\tilde{\Delta})|))$

Hodges-Lehmann estimator for the z-test: $\hat{\Delta} = \bar{Y} - \bar{X}$

Hodges-Lehmann estimator for Rank Sum Test:

$$\hat{\Delta} = \arg\min_{\tilde{\Delta} \in \mathbb{R}} \left| W_{obs}(\tilde{\Delta}) - \mu_W \right| = \arg\min_{\tilde{\Delta} \in \mathbb{R}} \left| W_{obs}(\tilde{\Delta}) - \frac{n(n+m+1)}{2} \right|$$

• Because from the 2-sided z-test, the probability in tails below or above W_{obs} (i.e. p-value) is monotone decreasing with $W_{obs}(\tilde{\Delta})$ for $W_{obs}(\tilde{\Delta}) \geq \mu_W$, and is monotone increasing with $W_{obs}(\tilde{\Delta})$ for $W_{obs}(\tilde{\Delta}) \leq \mu_W$. And p-value $= 1$ if $W_{obs}(\tilde{\Delta}) = \mu_W$.
• Robust against errors in the data. (see Example 1.19-1.20 at P20-21 on the notes.)

Median estimator: $\hat{\Delta} = \text{median}\{Y_j - X_i, 1 \leq i \leq n, 1 \leq j \leq m\}$ is unbiased.

1.2.7 Confidence Intervals

Lower Tail $1 - \alpha$ CI for location Δ from Wilcoxon Rank Sum Test: for $0 \leq \alpha \leq 1$,

$$C_\alpha(x, y) = \left\{ \tilde{\Delta} : G_{n,m}\left(W_{obs,\tilde{\Delta}}\right) > \alpha \right\}$$

where,
- *test statistic* $W_{\tilde{\Delta}}$, and the observed $W_{obs,\tilde{\Delta}}$ with observation $(X = x, Y = y)$;
- $G_{m,n}$ is the CDF of $W(R(X, Y - \tilde{\Delta}))$.

Prop 1.10: $Pr(\Delta \in C_\alpha(X, Y)) = 1 - \alpha$, if $\alpha \in \mathcal{G}_{m,n} = \{G_{m,n}(w) : w \in \mathcal{W}_{m,n}\}$.
Proof: Note that $W(R(X, Y - \tilde{\Delta}))$ is discrete, hence $G_{m,n}$ is a step-wise monotonically increasing function, where $G_{m,n}(w) > G_{m,n}(w')$ whenever $w > w'$.(i.e. Steps 1-1 corresponds to the discrete

w values.)
Further, $G_{m,n} : \mathcal{W}_{m,n} \to \mathcal{G}_{m,n}$ is invertible.
$\implies \Pr(G_{m,n}(W) \leq c) = c, \forall W \sim G_{m,n}, c \in \mathcal{G}_{m,n}$.
Also note, $\Pr(\Delta \in C_\alpha(X,Y)) = \Pr(G_{m,n}(W_{obs,\tilde{\Delta}}) > \alpha) = 1 - \Pr(G_{m,n}(W_{obs,\tilde{\Delta}}) \leq \alpha) = 1 - \alpha$ [EOP]

Prop 1.11: $Pr(\Delta \in C_\alpha(X,Y)) > 1-\alpha$, if $\alpha \notin \mathcal{G}_{m,n}$.
Proof: If $\alpha \notin \mathcal{G}_{m,n}$, consider $\alpha^- = \max\{g \in \mathcal{G}_{n,m} : g \leq \alpha\}$ is the largest value in $\mathcal{G}_{m,n}$ that is smaller than α. Then, $G_{m,n}(W_{obs,\tilde{\Delta}}) > \alpha$ is satisfied by any $\tilde{\Delta}$ that satisfies $G_{m,n}(W_{obs,\tilde{\Delta}}) > \alpha^-$.
$\Pr(\Delta \in C_\alpha(X,Y)) = \Pr(G_{m,n}(W_{obs,\tilde{\Delta}}) > \alpha) = \Pr(G_{m,n}(W_{obs,\tilde{\Delta}}) > \alpha^-) = 1 - \alpha^- > 1-\alpha$,
as $\alpha^- < \alpha$. [EOP]

Prop 1.12:
(i) $\tilde{\Delta} \uparrow \implies y - \tilde{\Delta} \downarrow \implies R_Y \downarrow \implies W_{obs,\tilde{\Delta}} \downarrow \implies G_{m,n}(W_{obs,\tilde{\Delta}}) \downarrow$ until hit the lower bound α;
(ii) $\tilde{\Delta}$ is large negative $\implies y - \tilde{\Delta} \uparrow \implies G_{m,n}(W_{obs,\tilde{\Delta}}) \uparrow 1$ when $\min_{j=1,\cdots,n} y_j - \tilde{\Delta} > \max_{i=1,\cdots,m} x_i$.

• Wilcoxon Rank Sum Test CI is robust against errors in data, as opposed to a simple t-test (which is quite sensitive).

1.3 Wilcoxon Signed Rank Test

Wilcoxon Signed Rank Test: a test for one-sample location problem using median,
(i) *Assumption*: $(X_i)_{i=1}^n \overset{iid}{\sim} F_X$, where F_X is continuous and symmetric about its median μ;
(ii) *Hypothesis*: $H_0 : \mu = \mu_0$ VS $H_1 : \mu > / < / \neq \mu_0$;
(iii) *Test Statistics*:

$$W(I,R) = \sum_{i=1}^n I_i R_i \sim \sum_{i=1}^n i J_i$$

where,
- $R_i = \sum_{j=1}^n \mathbb{I}\{|X_j - \mu_0| \leq |X_i - \mu_0|\}$ $(i = 1, \cdots, n)$ is the rank of $|X_i - \mu_0|$ in $\{|X_j - \mu_0|\}_{j=1}^n$;
- $I_i = \mathbb{I}\{X_i > \mu_0\}$ indicates whether X_i is above the median ;
- Under H_0, $I_i \sim Bern(1/2)$ and $R \sim U(\mathcal{P}_N)$, hence, $J \sim U\{0,1\}^n$.

• An *alternative form* of the test statistic $W(I', R) = W(2I-1, R)$ where $I' = sign(X - \mu_0)$, hence so-called "Signed Rank Test".
(iv) *Test Outcomes*: for observation $X = x$ and $W_{obs} = W(I(x), R(|x-\mu_0|))$,
- **positive shift** $H_1 : \mu > \mu_0$, compute $p^+ = \Pr(W(I,R) \geq W_{obs})$ and reject H_0 if $p^+ \leq \alpha$;
- **negative shift** $H_1 : \mu < \mu_0$, compute $p^- = \Pr(W(I,R) \leq W_{obs})$ and reject H_0 if $p^- \leq \alpha$;
- **non-zero shift** $H_1 : \mu \neq \mu_0$, compute p^+, p^- and reject H_0 if $2\min(p^-, p^+) \leq \alpha$ as the distribution of W is symmetric about $n(n+1)/4$.

• Large W value indicating $\mu_0 < \mu$(true).

Monte Carlo Permutation Test for Signed Rank Test: applies similar to Algorithm 1.

1.3.1 Normal Approximation

: Under H_0 the test statistic has,
- $\mu_W = E(W) = E\left(\sum_{i=1}^n I_i R_i\right) = \frac{1}{2} E\left(\sum_{i=1}^n R_i\right) = \frac{1}{2}\sum_{i=1}^n i = \frac{1}{2}\frac{n(n+1)}{2} = \frac{n(n+1)}{4}$;
- $V_W = Var(W) = Var(I)\sum_{i=1}^n i^2 = \frac{1}{4}\frac{n(n+1)(2n+1)}{6} = \frac{n(n+1)(2n+1)}{24}$.

Hence by CLT, $\frac{W - \mu_W}{\sqrt{V_W}} \sim N(0,1) \implies$ reject H_0 if $\left|\frac{W-\mu_W}{\sqrt{V_W}}\right| \geq z_{1-\alpha/2}$.

1.3.2 Test for Paired Data

Consider paired data $\{Y_i, Z_i\}_{i=1}^n$ sampled before and after a "treatment", hence Y, Z are very likely correlated within pair and independent between pairs.
• Rank Sum Test fails as joint independence is violated.

Prop 1.30: If Y, Z correlated but exchangeable, then $X = Y - Z$ is symmetric about 0.
Proof: Exchangeability of $Y, Z \implies f_{Y,Z}(y, z) = f_{Y,Z}(z, y)$. Then,

$$\begin{aligned} Pr(Z - Y \leq c) = Pr(Z \leq Y + c) &= \int_{y \in \mathcal{Y}} \int_{-\infty}^{y+c} f_{Y,Z}(y, z) dz dy \\ &= \int_{y \in \mathcal{Y}} \int_{-\infty}^{y+c} f_{Y,Z}(z, y) dz dy = \int_{z \in \mathcal{Z}} \int_{-\infty}^{z+c} f_{Y,Z}(y, z) dy dz \\ &= Pr(Y \leq Z + c) = Pr(Y - Z \leq c) \\ \implies P(Y - Z \leq c) = Pr(Y - Z \geq -c) \quad & \text{[EOP]} \end{aligned}$$

Lemma 1.31: By Prop 1.30, if Z and $Y - \Delta$ are correlated but exchangeable then $X = Y - Z$ is symmetric about $\mu = \Delta$.

• This satisfies the assumption of the Sign Rank Test, hence taking $H_0 : \mu = 0$, we can test for no location shift due to the treatment.
(see example 1.31 at P31 on the notes.)

1.3.3 Walsh Average and CI on the Median

Walsh Average (WA): $\left(\frac{X_i + X_j}{2}\right)_{i,j=1}^n$

Prop 1.32: the test statistic of the Wilcoxon Sign Rank Test can be expressed as the number of Walsh Averages exceeding the null μ_0,

$$W(I, R) = \sum_{i=1}^n I_i R_i = \sum_{i=1}^n \sum_{j=i}^n \mathbf{1}\{X_i + X_j > 2\mu_0\}$$

Proof: omitted, see PS1 solution of the optional question.

Hodges-Lehmann point estimate of the median from the Wilcoxon ranked sign test:

$$\hat{\mu} = median\left(\bigcup_{i=1}^n \bigcup_{j=i}^n \left\{\frac{X_i + X_j}{2}\right\}\right)$$

i.e. the median of the Walsh Averages.

CI for the Median in the 2-sided Test: for observation $X = x, W_{obs} = W(I(x), R(|x - \mu_0|))$ and significance level α,

$$\begin{aligned} C_\alpha &= \{\mu_0 : 2\min(p^+, p^-) > \alpha\} \\ &= \{\mu_0 : \min\{\Pr(W \geq W_{obs}(\mu_0) \mid H_0), \Pr(W \leq W_{obs}(\mu_0) \mid H_0)\} > \alpha/2\} \\ &\overset{(*)}{=} \{\mu_0 : \alpha/2 < G_n(W_{obs}(\mu_0)), G_n(W_{obs}(\mu_0) - 1) < 1 - \alpha/2\} \end{aligned}$$

where $\mathcal{W}_n = \{0, 1, \cdots, \frac{n(n+1)}{2}\}$ and $G_n(w) = \Pr(W \leq w | H_0), w \in \mathcal{W}_n$.

Prop 1.34: The $1 - \alpha$ CI is $C_\alpha = \{\mu : A_{(i_1)} < \mu < A_{(i_2)}\}$, with coverage $\Pr(\mu \in C_\alpha) = 1 - \alpha^-$, where
(i) $\{A_{(i)}\}_{i=1}^{n(n+1)/2}$ is Walsh averages sorted in increasing order;
(ii) $\alpha^- = 2 \max \{g \in \mathcal{G}_n : g \leq \alpha/2\}$ hence $\alpha > \alpha^-$;
(iii) $i_1 = w_{\alpha^-/2} + 1$ and $i_2 = \frac{n(n+1)}{2} - w_{\alpha^-/2}$;
(iv) $w_{\alpha^-/2}$ is the quantile of $W|H_0$ s.t. $G_n(w_{\alpha^-/2}) = {}^{\alpha^-}/_{2}$.
Proof:

$$\begin{aligned} C_\alpha &= \left\{\mu_0 : w_{\alpha^-/2} < W_{obs}\left(\mu_0\right), W_{obs}\left(\mu_0\right) - 1 < w_{1-\alpha^-/2}\right\}, \text{ from } (*) \text{ and Prop 1.34 (ii)} \\ &= \left\{\mu_0 : w_{\alpha^-/2} < \sum_{i=1}^{n} \mathbb{I}_{A_{(i)} > \mu_0} < w_{1-\alpha^-/2} + 1\right\} \\ &= \left\{\mu_0 : w_{\alpha^-/2} < \sum_{i} \mathbb{I}_{A_{(i)} > \mu_0} < n(n+1)/2 - w_{\alpha^-/2}\right\} \end{aligned}$$

To determine the bounds: note that $\mu_0 \uparrow \implies W_{obs}(\mu_0) \downarrow$, hence
- *lower bound*: by definition $\frac{\alpha^-}{2} < \frac{\alpha}{2}$ hence $w_{\alpha^-/2} + 1 \leq w_{\alpha/2} < W_{obs}(\mu_0)$;
- *upper bound*: Want to find the largest μ_0 can be is such that $G_n(W_{obs}(\mu_0)) > {}^{\alpha}/_{2} > {}^{\alpha^-}/_{2} \implies W_{obs}(\mu_0) = \sum_{i=1}^n \mathbb{I}_{A_{(i)} > \mu_0} > w_{\alpha^-/2}$. Note that if take $\mu_0 = A_{(i_2)}$, then $\sum_{i=1}^n \mathbb{I}_{A_{(i)} > \mu_0} = \frac{n(n+1)}{2} - i_2$. Hence $i_2 = \frac{n(n+1)}{2} - w_{\alpha^-/2}$ is the highest possible upper bounds, as $\forall \epsilon \in (0, A_{(i_2)} - A_{(i_2 - 1)})$, $W_{obs}(\mu_0) = \sum_{i=1}^n \mathbb{I}_{A_{(i)} > \mu_0} = w_{\alpha^-/2} + 1 > w_{\alpha^-/2}$. [EOP]

Prop 1.37: $W_{obs}(\mu_0) = \sum_{i=1}^n \mathbb{I}_{A_{(i)} > \mu_0}$.
Proof: Omitted, to be added.

• To get the equal-tail CI, sort the WA's and count the $w_{\alpha^-/2} + 1$-th from the top and the same counting up from the bottom.

2 Linear Smoothers

2.1 General Linear Smoothers

Model Setup: Given observations $(x_i, Y_i)_{i=1}^n$, we fit the model $Y_i = m(x_i) + \epsilon_i, i = 1, \cdots, n$, where,
(i) ϵ_i are zero-mean RVs s.t. $E(Y_i) = m(x_i)$;
(ii) the fitted model $\hat{Y}_i = \hat{m}(x_i)$ is a **smoother**.

Linear smoother:

$$\hat{Y} = \hat{m}(x) = \sum_{j=1}^{n} \ell_j(x; \mathbf{x}) Y_j = SY$$

with,
- $\ell(x) = (\ell_1(x; \mathbf{x}), \cdots, \ell_n(x; \mathbf{x}))^T$ a normalized vector s.t. $\sum_{j=1}^n \ell_j(x; \mathbf{x}) = 1$
- $S_{i,j} = \ell_j(x_i; \mathbf{x})$ s.t. $\hat{Y}_i = \sum_{j=1}^n S_{i,j} Y_j$. (row sum of S equals 1.)

2.1.1 Standard Linear Regression

Hat matrix: $S = X\left(X^TX\right)^{-1}X^T$.
Consider polynomial regression with a quadratic function: $Y_i = m(x_i) + \epsilon_i = \theta_1 + \theta_2 x_i + \theta_3 x_i^2 + \epsilon_i$.
Minimizing $RSS = \|Y - m(\mathbf{x})\|^2 = (Y - X\theta)^T(Y - X\theta) \implies \hat{\theta} = (X^TX)^{-1}X^TY$.
$\implies \hat{m}(x_i) = \hat{\theta}_1 + \hat{\theta}_2 x_i + \hat{\theta}_3 x_i^2 = (1, x, x^2)(X^TX)^{-1}X^TY = \sum_{j=1}^n \left[(1, x, x^2)\left(X^TX\right)^{-1}X^T\right]_j Y_j = \sum_{j=1}^n \ell_j(x; \mathbf{x})Y_j$

2.1.2 Degrees of freedom of a linear smoother

Prop 2.7: For a random vector v and a commensurate matrix A,

$$E\left(v^TAv\right) = E(v)^TAE(v) + \operatorname{tr}(A\operatorname{var}(v))$$

Proof:

$$\begin{aligned} E\left(v^TAv\right) &= E\left(tr(v^TAv)\right) = tr\left(AE(vv^T)\right), \text{ by equivalent permutation of trace} \\ &= tr\left(A\left(Var(v) + E(v)E(v)^T\right)\right), \text{ by definition of variance} \\ &= E(v)^TAE(v) + \operatorname{tr}(A\operatorname{var}(v)) \quad \text{[EOP]} \end{aligned}$$

Prop 2.8: Suppose variance-covariance matrix of Y is $\sigma^2 I_{n\times n}$ and smoother $\hat{m}(x) = SY$ with smoothing matrix S. If $RSS = \|Y - SY\|^2$, then

$$E(RSS) = \|(I - S)m(\mathbf{x})\|^2 + \sigma^2\left(n - df_{var}\right)$$

where expectation is taken over $Y|\mathbf{x}$ and $df_{\text{var}} = 2\operatorname{tr}(S) - \operatorname{tr}\left(S^TS\right)$ is the **degrees of freedom for the variance**.
Proof:

$$\begin{aligned} E(RSS) = E\left(\|(I-S)Y\|^2\right) &= E\left(Y^T(I-S)^T(I-S)Y\right) \quad \because \|Y - SY\| = \|(I-S)Y\| \\ &\text{by Prop 2.7 with } v = Y \text{ and } A = (I-S)^T(I-S) \\ &= m(\mathbf{x})^T(I-S)^T(I-S)m(\mathbf{x}) + \sigma^2\operatorname{tr}\left((I-S)^T(I-S)\right) \quad \because E(Y) = m(\mathbf{x}) \text{ and } Var(Y) = \sigma^2 I \\ &= \|(I-S)m(\mathbf{x})\|^2 + \sigma^2\left(\operatorname{tr}\left(S^TS\right) - 2\operatorname{tr}(S) + n\right) \\ &= \|(I-S)m(\mathbf{x})\|^2 + \sigma^2\left(n - df_{var}\right) \quad \text{[EOP]} \end{aligned}$$

- $df_{var} \uparrow \implies$ smoother complexity $\uparrow \implies E(RSS) \downarrow \implies$ fit to data improves.

DOF and variance for simple linear regression:
- $df_{\text{var}} = 2\operatorname{tr}(S) - \operatorname{tr}\left(S^TS\right) = p$, as $tr(S) = tr(S^TS) = p$ the number of parameters. (The hat matrix is essentially the identity matrix);
- $\hat{\sigma}^2 = \frac{RSS}{n-p}$, as $SX = X \implies (I - S)X\theta = (I - S)m(\mathbf{x}) = 0$.

Bias of the smoother: $E(\hat{Y} - m(\mathbf{x})) = (S - I)m(\mathbf{x})$

Estimator of the variance: $\hat{\sigma}^2 \approx \frac{RSS}{n - df_{var}}$, when bias is assumed to be small.

Mean Square Error (MSE): $\text{MSE}(\hat{m}(x)) = E\left((\hat{m}(x) - m(x))^2\right)$

Summed MSE (MSSE):

$$\begin{aligned}\text{MSSE}(\hat{m}(x)) &= \sum_{i=1}^{n} E\left((\hat{m}(x_i) - m(x_i))^2\right) \\ &= \|(I-S)m(\mathbf{x})\|^2 + \sigma^2 \operatorname{tr}\left(S^T S\right), \text{ if } Var(Y) = \sigma^2 I\end{aligned}$$

2.2 Kernel estimators and local linear regression smoothing

2.2.1 Local regression estimators: Nadaraya-Watson kernel estimator

Kernel $K(x)$: a continuous bounded symmetric probability density satisfying:
(i) $K(x) \geq 0$; (ii) $K(x) = K(-x)$; (iii) $\exists M < \infty : K(x) \leq M \forall x \in \mathbb{R}$;
(iv) $\inf K(x)dx = 1$, (v) $\lim_{|x|\to\infty} xK(x) = 0$.

- Epanechnikov Kernel: $K(x) = \frac{3\left(1-x^2\right)\mathbb{I}(|x|<1)}{4}$
- Normal Kernel: $K(x) = \frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}}$

Bandwidth $h > 0$: $w_j(x) = K\left(\frac{x-x_j}{h}\right)$ with $w_{i,j} = w_j(x_i)$.
•h sets the smoothing scale, and $|x - x_j| >> h \implies w_j(x) \to 0$.
• Large h means smoother line, hence likely under-fitting, vice versa.

Nadaraya-Watson smoother: with $\ell_j(x) = \frac{w_j(x)}{\sum_{k=1}^n w_k(x)}, j = 1, \ldots, n$

$$\hat{m}(x) = \sum_{j=1}^{n} \ell_j(x) Y_j$$

- a local linear kernel estimator of the linear smoother, and $S_{i,j} = \ell_j(x_i)$.

Prop 2.14: Nadaraya-Watson smoother $\hat{m}(x)$ solves the local linear regression

$$\hat{m}(x) = \arg\min_{a\in\mathbb{R}} \sum_{i=1}^{n} w_i(x) \left(Y_i - a\right)^2 = \sum_{j=1}^{n} \frac{w_j(x)}{\sum_{k=1}^n w_k(x)} Y_j$$

Proof: by solving $\frac{dC}{da} = 0$ where $C(a, x) = \sum_{i=1}^n w_i(x) \left(Y_i - a\right)^2$.

2.2.2 Higher order local linear regression estimators

Prop 2.15: Let X be $n \times p$ design matrix and $Y \sim N(X\theta, \sigma)$ with $\sigma^{-1} = W$, then:
(i) least squares estimator (LSE): $\hat{\theta} = (X^T W X)^{-1} X^T W Y$, and;
(ii) $\hat{Y} = SY$ with $S = X(X^T W X)^{-1} X^T W$.
Proof: Factorize $W = L^T L$, then $\hat{\theta} = \arg\min_\theta (Y - X\theta)^T W (Y - X\theta) = \arg\min_\theta \|LY - LX\theta\|^2 = \left((LX)^T LX\right)^{-1} (LX)^T (LY) = (X^T W X)^{-1} X^T W Y$ [EOP].

Prop 2.16 (local linear smoother): Consider $X_x = [X_{x,i}]_{i=1}^n$ be a $n \times 2$ design matrix with rows $X_{i,1} = (1, x_i - x)$. For $x \in \mathbb{R}, \theta_x = (\theta_{x,1}, \theta_{x,2})$, the *least-squares estimator for the effects θ_x fitted at x is:*

$$\hat{\theta}_x = \arg\min_{\theta_x \in \mathbb{R}^2} \sum_{i=1}^{n} \left(y_i - \theta_{x,1} - \theta_{x,2}\left(x_i - x\right)\right)^2 K\left(\left(x_i - x\right)/h\right)$$

Let $W_x = \operatorname{diag}(w_1(x), \cdots, w_n(x))$, the smoother:

$$\hat{m}(x) = (1,0)P_xY = (1,0)\left(X_x^TW_xX_x\right)^{-1}X_x^TW_xY$$

is a *local linear smoother* which gives the least-squares fit to the weighted linear regression at x. And the smoothing matrix S has entries $S_{i,j} = [P_{x_i}]_{1,j}, i,j = 1, \cdots, n$.
Proof: The least square problem can be rewritten as:

$$\hat{\theta}_x = \arg\min_{\theta_x} \left(Y - X_x\theta_x\right)^T W_x \left(Y - X_x\theta_x\right)$$

By **Prop 2.15**, $\hat{\theta}_x = P_xY = (X_x^TW_xX_x)^{-1}X_x^TW_xY$.
Note the fitted line as a function of z interpolating the data around x: $\hat{Y}_{x,z} = \hat{\theta}_{x,1} + \hat{\theta}_{x,2}(z-x)$.
Hence at x the fitted value is $\hat{m}(x) = \hat{Y}_{x,x} = \hat{\theta}_{x,1} = (1,0)P_xY$.
Let top row $\ell(x) = [P_x]_{1,:}^T = ([P_x]_{1,1}, \cdots, [P_x]_{1,n})^T$, then $\hat{Y}_i = \sum_{j=1}^n [P_{x_i}]_{1,j}Y_j$.

Prop 2.18 (local polynomial regression smoother of order p): Consider $X_x = [X_{x,i}]_{i=1}^n$ be the $n \times p+1$ design matrix with rows $X_{x,i} = \left(1, x_i - x, \frac{(x_i-x)^2}{2}, \ldots, \frac{(x_i-x)^p}{p!}\right)$ and the parameter that solves the least square problem $\hat{\theta}_x = \arg\min_{\theta_x \in \mathbb{R}^{p+1}} \sum_{i=1}^n \left(y_i - X_{x,i}\theta_x\right)^2 K\left(\frac{(x_i-x)}{h}\right)$.
The *local polynomial regression smoother*: $\hat{m}(x) = e_1P_xY$, with $e_1 = (1, 0, \cdots, 0)$ a $1 \times p+1$ vector.

Prop 2.19: The local polynomial regression smoother is a Nadaraya-Watson smoother, $\hat{m}(x) = \sum_{j=1}^n \ell_j(x)Y_j$, with $\sum_{j=1}^n \ell_j(x) = \sum_{j=1}^n [P_x]_{1,j} = 1$.
Proof: (smoothing constant data to constant function.) Suppose constant data $Y_i = c, i = 1, \cdots, n$. Then the LSE is: $\hat{\theta}_x = (c, 0, \cdots, 0)^T$. Note that the fitted smoother is:

$$\begin{aligned}
\hat{m}(x) &= \hat{\theta}_{x,1} = c \\
\hat{m}(x) &= \ell(x)^TY = \sum_{j=1}^n \ell_j(x)c \\
\implies \sum_{j=1}^n \ell_j(x) &= 1 \quad \text{[EOP]}
\end{aligned}$$

2.2.3 Bias

Prop 2.21: Consider local polynomial regression smoothing at order p, $\forall a \in \mathbb{R}^p, x \in \mathbb{R}$

$$\sum_j \ell_j(x)\left[a_1\left(x_j - x\right) + \ldots + \frac{a_p}{p!}\left(x_j - x\right)^p\right] = 0$$

Proof: Suppose $m(x)$ is a polynomial of order p: (the Taylor Series about x evaluated at x_j ends at order p)

$$m\left(x_j\right) = m(x) + m^{(1)}(x)\left(x_j - x\right) + \ldots + \frac{m^{(p)}(x)}{p!}\left(x_j - x\right)^p$$

and observe $m(x)$ interpolates $p+1$ data points, i.e. $Y_i = m(x_i)$. Hence least-square smoother is $\hat{m}(x) = m(x)$. Taking expectation on both sides gives: (by **Prop 2.19**)

$$E[\hat{m}(x)] = \textstyle\sum_{j=1}^n \ell_j(x)\underbrace{E[Y_j]}_{m(x_j)} = \sum_{j=1}^n \ell_j(x)m(x) = E[m(x)] = m(x) \implies \sum_j \ell_j(x)\left(m\left(x_j\right) - m(x)\right) = 0$$

$$\implies \textstyle\sum_{j=1}^n \ell_j(x)\left[m^{(1)}(x)\left(x_j - x\right) + \ldots + \frac{m^{(p)}(x)}{p!}\left(x_j - x\right)^p\right] = 0$$

For all $a \in \mathbb{R}^p, x \in \mathbb{R}$, choose $m(x)$ to be a polynomial of order p matching $m^{(k)}(x) = a_k, k = 1, \cdots, p$. [EOP]

Prop 2.22: If smoother $m(x)$ has at least $p+1$ continuous derivatives at every $x \in \mathbb{R}$, then the bias of a local polynomial regression smoother of order p is: (for some ξ_j between x and x_j)

$$E(\hat{m}(x)) - m(x) = \sum_{j=1}^{n} \frac{m^{(p+1)}(\xi_j)}{(p+1)!} (x_j - x)^{p+1} \ell_j(x)$$

Proof: Suppose $m(x)$ is a polynomial of infinite order, then the Taylor series has a remainder:

$$\begin{aligned} E(\hat{m}(x)) - m(x) &= \sum_{j=1}^{n} \ell_j(x) m(x_j) - m(x) \\ &= \sum_{j=1}^{n} \ell_j(x) \left[m(x) + m^{(1)}(x)(x_j - x) + \ldots + \frac{m^{(p)}(x)}{p!}(x_j - x)^p + \frac{m^{(p+1)}(\xi_j)}{(p+1)!}(x_j - x)^{p+1} \right] - m(x) \\ &= \sum_{j=1}^{n} \ell_j(x) \frac{m^{(p+1)}(\xi_j)}{(p+1)!}(x_j - x)^{p+1}, \text{ by } \textbf{Prop 2.21} \text{ and } \textbf{Prop 2.19} \text{ [EOP]} \end{aligned}$$

(Some remarks omitted on P45.)

2.3 Choice of Bandwidth h

2.3.1 Risks: MSE

We aim to choose the bandwidth to minimize the mean square error (MSE) of the fit:

$$\begin{aligned} \text{MSE}(h, x) &= E\left\{ \left(Y' - \hat{m}_h(x)\right)^2 \right\} = E\left\{ \left(\epsilon' + m(x) - \hat{m}_h(x)\right)^2 \right\} \\ &= \operatorname{var}(\epsilon') + E\left\{ (m(x) - \hat{m}_h(x) \right\}^2 \\ &= \operatorname{var}(Y') + \text{MSE}(\hat{m}_h(x)) = \underbrace{Var(Y')}_{Noise} + \underbrace{(E[\hat{m}(x) - m(x)])^2}_{Bias^2} + \underbrace{E[(\hat{m}(x) - E[\hat{m}(x)])^2]}_{Variance} \end{aligned}$$

$\bullet h \downarrow 0 \implies$ Bias ↓ and Variance ↑.

2.3.2 Cross-Validation (CV)

Mean Summed Square Error (MSSE) for the fit:

$$\text{MSSE}(\hat{m}_h(\mathbf{x})) = \frac{1}{n} \sum_i \text{MSE}(\hat{m}_h(x_i)) = \sigma^2 + \frac{1}{n} \sum_i E[(Y_i' - \hat{m}_h(x_i))^2] = \sigma^2 + \text{MSSE}(h)$$

Leave-one-out cross validation (LOOCV):

$$h^* = \arg\min_h \widehat{MSSE}(h) = \arg\min_h \frac{1}{n} \sum_i E[(Y_i' - \hat{m}_h^{(-i)}(x_i))^2]$$

Prop 2.30 (LOOCV for Linear Smoothers): Let $\hat{m}(x) = \ell(x; \mathbf{x})^T Y$ be a linear smoother with $\ell(x; \mathbf{x}) = (\ell_1(x; \mathbf{x}), \ldots, \ell_n(x; \mathbf{x}))^T$.

If $\ell_j\left(x_i;\mathbf{x}_{-i}\right)=\frac{\ell_j(x_i;\mathbf{x})}{\sum_{k\neq i}\ell_k(x_i;\mathbf{x})}$ then $\widehat{MSSE}(h)=n^{-1}\sum_{i=1}^n\left(\frac{Y_i-\hat{m}_h(x_i)}{1-S_{i,i}}\right)^2$,
where $S_{i,j}=\ell_j\left(x_i;\mathbf{x}\right), i,j=1,\ldots,n$ is our earlier smoothing matrix notation (so $\hat{y}=SY$ etc).
Proof: For each $i=1,\cdots,n$,

$$\begin{aligned}
Y_i-\hat{m}_h^{(-i)}\left(x_i\right) &= Y_i-\sum_{j\neq i}\ell_j\left(x_i;\mathbf{x}_{-i}\right)Y_j = Y_i-\sum_{j\neq i}\frac{\ell_j\left(x_i;\mathbf{x}\right)}{\sum_{k\neq i}\ell_k\left(x_i;\mathbf{x}\right)}\\
&= Y_i-\sum_{j\neq i}\frac{S_{i,j}}{\sum_{k\neq i}S_{i,k}}Y_j = Y_i-\sum_{j\neq i}\frac{S_{i,j}}{1-S_{i,i}}Y_j,\text{ since }\sum_k S_{i,k}=1\\
&= \frac{1}{1-S_{i,i}}\left(Y_i-\sum_j S_{i,j}Y_j\right)=\frac{1}{1-S_{i,i}}\left(Y_i-\hat{m}_h\left(x_i\right)\right)\quad\text{[EOP]}
\end{aligned}$$

Generalized Cross-Validation (GCV): $GCV(h)=n^{-1}\sum_{i=1}^n\left(\frac{Y_i-\hat{m}_h(x_i)}{1-\nu/n}\right)^2$,
where ν is the degree of freedom, which can be measure by:
- $\nu=tr(S)=\sum_i S_{i,i}$, or;
- $\nu=2tr(S)-tr(S^TS)$ (effective DOF).

2.4 Penalized Regression

Roughness penalty: For $f\in C^2[a,b]$, the loss function is
$L(f)=RSS(f)+\lambda J(f)=\sum_{i=1}^n(y_i-f(x_i))^2+\int_a^b(f''(x)^2dx$.

2.4.1 Splines

For the knots $a=\xi_0<\xi_1<\cdots<\xi_n<\xi_{n+1}=b$,
Cubic Spline: g is a cubic polynomial over $(\xi_i,\xi_{i+1}), i=1,\cdots,n$ and has continuous first and second derivatives at the knots ξ_i's.
Natural Splines: a cubic spline that is linear on the tails ($[a,\xi_1]\cup[\xi_n,b]$).
M-th Order Spline: a piece-wise $M-1$ degree polynomial with $M-2$ continuous derivatives at the knots.
Smoothing Spline: $f^*=\arg\min_f L(f)=\arg\min_f RSS(f)+\lambda J(f)$.
Prop 2.37: A smoothing spline is a natural cubic spline.
Proof: By contradiction, suppose $L(f)$ is minimized by h which is not a natural cubic spline. Hence $L(h)<L(g),\forall g$ even if g is a natural cubic spline. Suppose g is interpolating the data, then $g(x_i)=h(x_i),\forall i$ and hence $RSS(g)=RSS(h)$ matches. But by Prop 2.41, $J(g)=\int_a^b\left(g''(x)\right)^2dx\leq\int_a^b\left(h''(x)\right)^2dx=J(h)$, resulting $L(g)\leq L(h)$ a contradiction. [EOP]

Interpolating Natural Cubic Spline: $g:[a,b]\to\mathbb{R}$ satisfying,
(i) $g(x_i)=f_i, i=1,\cdots,n$; (ii) cubic on $[x_k,x_{k+1}]$ and linear on $[a,x_1]\cup[x_n,b]$;
(iii) $g\in C^2[a,b]$; (iv) $g''(x_1)=g''(x_n)=0$.

Prop 2.41: The interpolating natural cubic spline g has the smallest roughness penalty among all the interpolating splines, i.e. $\int_a^b\left(g''(x)\right)^2dx\leqslant\int_a^b\left(h''(x)\right)^2dx,\forall h\in C^2[a,b]$.
Proof: Since $g''(x)=0\forall x\in[a,x_1]\cup[x_n,b]$ (linear on tails), sufficient to show: $\int_{x_1}^{x_n}\left(g''(x)\right)^2dx\leqslant\int_{x_1}^{x_n}\left(h''(x)\right)^2dx$.

First claim: $\int_{x_1}^{x_n} h''(x)g''(x)dx = \int_{x_1}^{x_n} \left(g''(x)\right)^2 dx$. (to be shown later.)

$$\begin{aligned}0 &\leqslant \int_{x_1}^{x_n} \left(h''(x) - g''(x)\right)^2 dx \\ &= \int_{x_1}^{x_n} \left(h''(x)\right)^2 dx - 2\int_{x_1}^{x_n} h''(x)g''(x)dx + \int_{x_1}^{x_n} \left(g''(x)\right)^2 dx \\ &= \int_{x_1}^{x_n} \left(h''(x)\right)^2 dx - \int_{x_1}^{x_n} \left(g''(x)\right)^2 dx, \text{ by the claim.}\end{aligned}$$

The claim is true because:

$$\begin{aligned}&\int_{x_1}^{x_n} g''(x)\left(h''(x) - g''(x)\right) dx = \left[g''(x)\left(h'(x) - g'(x)\right)\right]_{x_1}^{x_n} - \int_{x_1}^{x_n} g'''(x)\left(h'(x) - g'(x)\right) dx \\ &\quad = 0 - \int_{x_1}^{x_n} g'''(x)\left(h'(x) - g'(x)\right) dx \quad \because g''(x_1) = g''(x_n) = 0 \\ &\quad \sum_{k=1}^{n-1} \int_{x_k}^{x_{k+1}} g'''(x)\left(h'(x) - g'(x)\right) dx = c_k \int_{x_k}^{x_{k+1}} \left(h'(x) - g'(x)\right) dx \\ &\quad = \sum_{k=1}^{n-1} c_k \left(h\left(x_{k+1}\right) - h\left(x_k\right) - g\left(x_{k+1}\right) + g\left(x_k\right)\right) = 0 \quad \because h(x_k) = g(x_k) \forall k \text{ [EOP]}\end{aligned}$$

B-Splines: For knots $a = \xi_0 < \xi_1 < \cdots < \xi_n < \xi_{n+1} = b$, define new knots on both tails,
•$\tau_0 \leq \tau_1 \leq \ldots \leq \tau_M = \xi_0 = a$, ($M$ intervals below)
•$\tau_{j+M} = \xi_j, j = 0, \ldots, n+1$
•$b = \xi_{n+1} = \tau_{n+M+1} \leq \tau_{n+M+2} \leq \ldots \leq \tau_{n+2M+1}$(and M above) .
Then the basis for the M-th order spline is defined recursively as:

$$B_{i,m}(x) = \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(x), \quad x \in [\tau_i, \tau_{i+m}]$$

with $B_{i,1} = \begin{cases} 1, \tau_i \leq x < \tau_{i+1} \\ 0, o/w \end{cases}$.

Prop 2.42: For some $\beta \in \mathbb{R}^{n+4}$, $f(x;\beta) = \sum_{j=1}^{n+4} B_j(x)\beta_j$ is a cubic spline, with $B_j(x) = B_{j,M}(x)$. For any cubic spline function g, $\exists \beta \in \mathbb{R}^{n+4}$ s.t. $f = g$.
•$n+4$ because, for a cubic spline, $M = 3$ hence $\tau_3 = a$. The first non-zero spline is $B_{1,M}$ with support $[\tau_1, \tau_{1+M}]$, as the previous on $B_{0,M}$ has support $[\tau_0, \tau_M = a]$ (not overlapping with $[a,b]$). The last non-zero spline is $B_{n+M,M}$ with support $[\tau_{n+M}, \tau_{n+M+1} = b]$, as the next one $B_{n+M+1,M}$ has support $[\tau_{n+M+1} = b, \tau_{n+M+2}]$ (not overlapping with $[a,b]$).

Spline Smoother with knots at the data: $\hat{m}(x) = \sum_{j=1}^{n+4} \hat{\beta}_j B_j(x)$, with $\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^{n+4}} C(\beta)$, where $C(\beta) = \sum_{i=1}^{n} \left(y_i - f\left(x_i;\beta\right)\right)^2 + \lambda \int_a^b \left(f''(x;\beta)\right)^2 dx$
• The spline smoother is a smoothing spline and a natural cubic spline.

Prop 2.47: Suppose data $y = (y_{1:n})$, B a $n \times (n+4)$ matrix with entries $\{B_{i,j}\}_{i=1:n}^{j=1:n+4} = \{B_j(x_i)\}_{i=1:n}^{j=1:n+4}$, and $\Omega_{j,k} = \int_a^b B_j''(x)B_k''(x)dx$, for $j,k = 1, \ldots, n+4$. Then:
(i) *estimated regression coefficients*: $\hat{\beta} = (B^T B + \lambda\Omega B)^{-1} B^T y$, and;
(ii) *fitted values*: $\hat{y} = Sy = B\hat{\beta}$ with $S = B(B^T B + \lambda\Omega B)^{-1} B^T$.

Proof: minimizing the loss function:

$$\begin{aligned}
C(\beta) &= \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{n+4} B_j\left(x_i\right)\beta_j\right)^2 + \lambda \sum_{j,j'=1}^{n+4} \beta_j \beta_{j'} \int_a^b B_j''(x) B_{j'}''(x) dx \\
&= (y - B\beta)^T (y - B\beta) + \lambda \beta^T \Omega \beta \\
\implies \frac{\partial C}{\partial \beta} &= -2B^T y + 2\lambda B^T B y + 2\lambda \Omega \beta = 0 \quad \text{[EOP]}
\end{aligned}$$

3 Bootstrap Samplers

3.1 Empirical Cumulative Distribution Function (ECDF)

ECDF: $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\left(X_i \leq x\right) = \frac{|\{i|X_i \leq x\}|}{n}$.

Properties of ECDF:
- *Unbiased*: $\mathbb{E}[F_n(x)] = F(x), \forall x \in \mathbb{R}$ and $n \geq 1$;
- *(Strongly) consistent*: $F_n(x) \overset{a.s}{\to} F(x), \forall x, n \to \infty$;
- *Asymptotically normal*: $\sqrt{n}\left(F_n(x) - F(x)\right) \overset{d}{\to} \mathcal{N}(0, F(x)(1 - F(x))), \forall x, n \to \infty$.

3.2 Monte Carlo Integration

Let $Y \in \mathbb{R}$ be a continuous or discrete random variable with cdf G, consider $\eta = \mathbb{E}(\phi(Y)) = \int_{\mathbb{R}} \phi(y) dG(y)$, where $\phi : \mathbb{R} \to \mathbb{R}$. Then,

Monte Carlo Estimator of Expectation η: $\widehat{\eta}_B = \frac{1}{B}\sum_{j=1}^{B} \phi\left(Y^{(j)}\right)$.
Properties of MC estimator:
- Unbiased: $E[\hat{\eta}_B] = \eta$;
- (Strongly) consistent: $\hat{\eta}_B \overset{a.s.}{\to} \eta$ as $B \to \infty$;
- Asymptotically normal: $\sqrt{B}(\hat{\eta}_B - \eta) \overset{d}{\to} N(0, \sigma^2 = \mathbb{V}(\phi(Y)))$ as $B \to \infty$.

3.3 Monte Carlo estimators of the mean and variance of Y

$$\begin{aligned}
\widehat{\mu}_B &= \frac{1}{B}\sum_{j=1}^{B} Y^{(j)} \overset{as}{\to} \mathbb{E}(Y), \\
\widehat{\sigma}_B^2 &= \frac{1}{B}\sum_{j=1}^{B}\left(Y^{(j)} - \widehat{\mu}_B\right)^2 = \frac{1}{B}\sum_{j=1}^{B}\left(Y^{(j)}\right)^2 - \left(\frac{1}{B}\sum_{j=1}^{B} Y^{(j)}\right)^2 \overset{as}{\to} \mathbb{V}(Y).
\end{aligned}$$

3.4 Confidence Intervals (CI)

3.4.1 Normal CI

$$
\begin{aligned}
\mathbb{P}\left(\hat{\theta}_n - z_{\alpha/2}\widehat{\sigma}_{n,B} \leq \theta \leq \hat{\theta}_n + z_{\alpha/2}\widehat{\sigma}_{n,B}\right) &= \mathbb{P}\left(\frac{\hat{\theta}_n - \theta}{\widehat{\sigma}_{n,B}} \in \left[-z_{\alpha/2}, +z_{\alpha/2}\right]\right) \\
&\simeq \mathbb{P}\left(\frac{\hat{\theta}_n - \theta}{\sqrt{\mathbb{V}_F\left(\hat{\theta}_n\right)}} \in \left[-z_{\alpha/2}, +z_{\alpha/2}\right]\right), \text{ by Bootstrap estimators of the mean and variance} \\
&\simeq 1 - \alpha, \text{ by CLT}
\end{aligned}
$$

3.4.2 $1-\alpha$ Bootstrap CI

$C_n^* = \left[2\hat{\theta}_n - \hat{q}^{\theta*}_{1-\alpha/2}, 2\hat{\theta}_n - \hat{q}^{\theta*}_{\alpha/2}\right]$, where $\hat{q}^{\theta*}_{\alpha/2}$ and $\hat{q}^{\theta*}_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ quantile of the bootstrap sample $\hat{\theta}_n^{*(1)}, \cdots, \hat{\theta}_n^{*(B)}$.

3.5 Bias Estimation

Bootstrap Bias Estimator: $\hat{b}_{n,B} = \left(\frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_n^{*(j)}\right) - \hat{\theta}_n$.

Bootstrap Bias Corrected Estimator for θ: $\hat{\theta}_{n,bcB} = \hat{\theta}_n - \hat{b}_{n,B} = 2\hat{\theta}_n - \left(\frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_n^{*(j)}\right)$.

3.6 Properties of Bootstrap

• Bootstrap uses the distribution of $R^* = \hat{\theta}_n^* - \hat{\theta}_n$ to approximate the unknown distribution $R_n = \hat{\theta}_n - \theta$.

Consistency of Bootstrap:
(i) Weak consistency: for some metric ρ of the CDF space $\rho\left(\widetilde{H}_n, \widetilde{H}^*_{F_n}\right) \xrightarrow{p} 0 \quad n \to \infty$;
(ii) Strong consistency: the convergence is almost surely (e.g. under Kolmogorov metric $K(G, L) = \sup_{x\in\mathbb{R}}|G(x) - L(x)|$, strong consistency is: $K\left(\widetilde{H}_n, \widetilde{H}^*_{F_n}\right) \xrightarrow{a.s.} 0$ as $n \to \infty$);
where:

$$
\begin{aligned}
\widetilde{H}_n(x) &:= \mathbb{P}_F\left(a_n R_n \leq x\right) \\
\widetilde{H}^*_{F_n}(x) &:= \mathbb{P}_{F_n}\left(a_n R_n^* \leq x\right)
\end{aligned}
$$

Implication of Bootstrap Consistency: both variance and bias can be consistently estimated,

$$
\begin{aligned}
&\tfrac{\mathbb{V}_{F_n}\left(\hat{\theta}_n^*\right)}{\mathbb{V}_F\left(\hat{\theta}_n\right)} \xrightarrow{p} 1 \quad n \to \infty \\
&\tfrac{\mathbb{E}_{F_n}\left(\hat{\theta}_n^*\right) - \hat{\theta}_n}{\mathbb{E}_F\left(\hat{\theta}_n\right) - \theta} \xrightarrow{p} 1 \quad n \to \infty
\end{aligned}
$$

3.7 Bootstrap for regression

Consider $Y_i = h(X_i, \theta) + \epsilon_i$. The fitted model has parameters $\hat{\theta}_n = t((X_1, Y_1), \cdots, (X_n, Y_n))$, with fitted residuals $\hat{\epsilon}_i = Y_i - h(X_i, \hat{\theta}_n)$

3.7.1 Nonparametric paired bootstrap

Algorithm 3: Non-parametric Paired Bootstrapping for the Variance Estimates

```
Initialize the size of bootstrap sample $B$.
for $b = 1, \cdots, B$ do
    Randomly sample $n$ response-predictor pairs $(X_i^{(b)}, Y_i^{(b)})_{i=1}^n$ from data with replacement.
    Fit the model and compute estimates of the parameters:
      $\hat{\theta}_n^{*(b)} = t((X_1^{*(b)}, Y_1^{*(b)}), \cdots, (X_n^{*(b)}, Y_n^{*(b)}))$.
end
Return: $\widehat{\sigma}_{n,B}^2 = \frac{1}{B}\sum_{j=1}^B \left(\hat{\theta}_n^{*(j)} - \frac{1}{B}\sum_{j=1}^B \hat{\theta}_n^{*(j)}\right)^2$
```

3.7.2 Semi-parametric paired bootstrap

Algorithm 4: Semi-parametric Paired Bootstrapping for the Variance Estimates

```
Initialize the size of bootstrap sample $B$.
for $b = 1, \cdots, B$ do
    Randomly sample $n$ fitted residuals $(\hat{\epsilon}_i^{*(b)})_{i=1}^n$ with replacement.
    for $i = 1, \cdots, n$ do
        Set $Y_i^{*(b)} = h\left(X_i, \hat{\theta}_n\right) + \hat{\varepsilon}_i^{*(b)}$
    end
    Refit the model and compute estimates of the parameters:
      $\hat{\theta}_n^{*(b)} = t((X_1, Y_1^{*(b)}), \cdots, (X_n, Y_n^{*(b)}))$.
end
Return: $\widehat{\sigma}_{n,B}^2 = \frac{1}{B}\sum_{j=1}^B \left(\hat{\theta}_n^{*(j)} - \frac{1}{B}\sum_{j=1}^B \hat{\theta}_n^{*(j)}\right)^2$
```

3.8 Studentized Bootstrap CI

Pivot: Suppose $X_{1:n}$ iid RV with pdf $f(x;\theta)$, then $Q_n = Q(X_{1:n};\theta)$ is a *pivotal quantity* for θ if it is a RV with distribution independent of θ.

Asymptotic pivot: $Q_n = \frac{\widehat{\theta}_n - \theta}{\sqrt{\widehat{V}_n(\hat{\theta}_n)}} \xrightarrow{d} \mathcal{N}(0,1)$

- If $X_{1:n} \overset{iid}{\sim} N(\mu, \sigma^2)$ with $\hat{\mu}_n = \frac{1}{n}\sum_i X_i$ and $\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_i (X_i - \hat{\mu}_n)^2$, then $Q_n = \frac{\widehat{\mu}_n - \mu}{\widehat{\sigma}_n/\sqrt{n}} \sim t_{(n-1)}$.

Algorithm 5: Semi-parametric Paired Bootstrapping for the Variance Estimates

```
Initialize the size of bootstrap sample $B$.
The observed parameter from the model: $\hat{\theta}_n$.
for $b = 1, \cdots, B$ do
    Randomly sample $n$ data points $(X_1^{(b)}, \cdots, X_n^{(b)})$ with replacement.
    Refit the model and compute estimates of the parameters $\hat{\theta}_n^{*(b)} = t(X_1^{(b)}, \cdots, X_n^{(b)})$ and
      the corresponding variance estimate $\widehat{V}_n^*(\hat{\theta}_n^{*(b)})$.
    Compute the asymptotic pivotal quantity: $Q_n^{*(b)} = \frac{\widehat{\theta}_n^{*(b)} - \widehat{\theta}_n}{\sqrt{\widehat{V}_n^*\left(\widehat{\theta}_n^{*(b)}\right)}}$
end
Return: $1 - \alpha$ studentized bootstrap confidence interval is:

    $$C_{ns}^* = \left[\widehat{\theta}_n - q_{1-a/2}^{Q*}\sqrt{\widehat{V}_n\left(\hat{\theta}_n\right)}, \widehat{\theta}_n - q_{a/2}^{Q*}\sqrt{\widehat{V}_n\left(\hat{\theta}_n\right)}\right]$$

 where $q_{a/2}^{Q*}$ and $q_{1-a/2}^{Q*}$ are the $\alpha/2$ and $1 - \alpha/2$ quantiles of the bootstrap sample
 $(Q_n^{*(1)}, \cdots, Q_n^{*(B)})$.
```

4 Hidden Markov Models

4.1 Discrete-state Hidden Markov models

4.1.1 Discrete Markov chain

Markov Chain (MC): a random process $X_{0:T}$ s.t. $\forall t \geq> 0$ and $x_{0:t+1} \in \mathcal{X}$ (the **state space**),

$$\mathbb{P}\left(X_{t+1} = x_{t+1} \mid X_t = x_t, \ldots, X_0 = x_0\right) = \mathbb{P}\left(X_{t+1} = x_{t+1} \mid X_t = x_t\right)$$

- **Homogeneous** MC is $A_{i,j} := \mathbb{P}\left(X_{t+1} = j \mid X_t = i\right) \perp t, \quad \forall i, j \in \mathcal{X}$.

Joint pmf Under Homogeneous MC:

$$p\left(x_{0:T}\right) := \mathbb{P}\left(X_0 = x_0, \ldots, X_T = x_T\right) = \mathbb{P}\left(X_0 = x_0\right) \prod_{t=1}^{T} \mathbb{P}\left(X_t = x_t \mid X_{t-1} = x_{t-1}\right) = \mu_{x_0} \prod_{t=1}^{T} A_{x_{t-1}, x_t}$$

4.1.2 Hidden Markov Model (HMM)

With the homogeneous MC $X_{0:T} \in \mathcal{X}$ and the $Y_{1:T} \in \mathcal{Y}$ (the **observation space**),

$$\mathbb{P}\left(Y_1 = y_1, \ldots, Y_T = y_T \mid X_0 = x_0, \ldots, X_T = x_T\right) = \prod_{t=1}^{T} \mathbb{P}\left(Y_t = y_t \mid X_t = x_t\right)$$

- **Homogeneous HMM**: if the **emission** $g_x(y) := \mathbb{P}\left(Y_t = y \mid X_t = x\right) \perp t, \forall t$.

Joint Probability of Hidden States and Observations:

$$\mathbb{P}\left(X_{0:T} = x_{0:T}, Y_{1:T} = y_{1:T}\right) = \mathbb{P}(Y_{1:T} = y_{1:T} | X_{0:T} = x_{0:T}) p(x_{0:T}) = \mu_{x_0} \prod_{t=1}^{T} g_{x_t}\left(y_t\right) A_{x_{t-1}, x_t}$$

4.1.3 Inference in HMM

(1) Forward Filtering: want to compute $p\left(x_t \mid y_{1:t}\right) = \frac{p(x_t, y_{1:t})}{p(y_{1:t})} = \frac{p(x_t, y_{1:t})}{\sum_{x'_t \in \mathcal{X}} p(x'_t, y_{1:t})}$

Algorithm 6: Forward $\alpha-$Recursion (To compute $p\left(x_t \mid y_{1:t}\right)$)

```
Initialize $\alpha_0(i) = \mu_i$ for $i \in \mathcal{X} = \{1, \cdots, K\}$.
for $t = 1, \cdots, T$ do
    for $x_t = j \in \mathcal{X}$ do
        $\alpha_t(j) = p\left(x_t, y_{1:t}\right) = \sum_{x_{t-1} \in \mathcal{X}} p\left(x_t, x_{t-1}, y_t, y_{1:t-1}\right)$
        $= \sum_{x_{t-1} \in \mathcal{X}} p\left(y_t \mid x_t, x_{t-1}, y_{1:t-1}\right) p\left(x_t \mid x_{t-1}, y_{1:t-1}\right) p\left(x_{t-1}, y_{1:t-1}\right)$
        $= p\left(y_t \mid x_t\right) \sum_{x_{t-1} \in \mathcal{X}} p\left(x_t \mid x_{t-1}\right) p\left(x_{t-1}, y_{1:t-1}\right) = g_j\left(y_t\right) \sum_{i=1}^{K} A_{i,j} \alpha_{t-1}(i)$
    end
end
```

- Thereby compute: **filter pmf**: $p\left(x_t \mid y_{1:t}\right) = \frac{p(x_t, y_{1:t})}{p(y_{1:t})} = \frac{\alpha_t(x_t)}{\sum_{x \in \mathcal{X}} \alpha_t(x)}$

(2) Likelihood: $p_{1:T} = \sum_{x \in \mathcal{X}} \alpha_T(x_t)$

(3) Prediction: $p_{x_t \mid y_{1:s}}, s < t$

- α_t may diminish for large t, an alternative is **predict-update recursion**:

$$p\left(x_t \mid y_{1:t-1}\right) = \sum_{x_{t-1} \in \mathcal{X}} p\left(x_t \mid x_{t-1}\right) p\left(x_{t-1} \mid y_{1:t-1}\right) \quad \text{(predict)}$$

$$p\left(x_t \mid y_{1:t}\right) = \frac{g_{x_t}\left(y_t\right) p\left(x_t \mid y_{1:t-1}\right)}{\sum_{x'_t \in \mathcal{X}} g_{x'_t}\left(y_t\right) p\left(x'_t \mid y_{1:t-1}\right)} \quad \text{(update)}$$

(4) Forward-backward Smoothing: want to compute $p\left(x_t \mid y_{1:T}\right) = \frac{p(x_t, y_{1:T})}{p(y_{1:T})} = \frac{p(x_t, y_{1:t}) p(y_{t+1:T} \mid x_t)}{p(y_{1:T})}$

- $p(x_t | y_{1:t})$ can be computed by forward recursion (Algorithm 6), and $p(y_{t+1:T} | x_t)$ can be computed by backward recursion.

Algorithm 7: Backward $\beta-$Recursion

```
Initialize $\beta_T(i) = 1$ for $i \in \mathcal{X} = \{1, \cdots, K\}$.
for $t = T, \cdots, 1$ do
    for $x_t = i \in \mathcal{X}$ do
        $$\begin{aligned}\beta_{t-1}(i) = p(y_{t:T} \mid x_{t-1}) &= \sum_{x_t \in \mathcal{X}} p(y_{t:T}, x_t \mid x_{t-1}) \\ &= \sum_{x_t \in \mathcal{X}} p(y_t \mid y_{t+1:T}, x_t, x_{t-1})\, p(y_{t+1:T}, x_t \mid x_{t-1}) \\ &= \sum_{x_t \in \mathcal{X}} p(y_t \mid x_t)\, p(y_{t+1:T} \mid x_t, x_{t-1})\, p(x_t \mid x_{t-1}) \\ &= \sum_{x_t \in \mathcal{X}} p(y_t \mid x_t)\, p(y_{t+1:T} \mid x_t)\, p(x_t \mid x_{t-1}) = \sum_{j=1}^{K} g_j(y_t) A_{i,j} \beta_t(j)\end{aligned}$$
    end
end
```

• Thereby compute **smoothing pmf**: $p(x_t \mid y_{1:T}) = \frac{p(x_t, y_{1:T})}{p(y_{1:T})} = \frac{\alpha_t(x_t)\beta_t(x_t)}{\sum_{x \in \mathcal{X}} \alpha_t(x)\beta_t(x)}$

- where $\alpha_t(x_t)$ comes from the forward recursion.

• Computational complexity: $O(T|\mathcal{X}|^2)$, because:

(i) iterate T times;

(ii) iterate (twice) over x_t and x_{t-1} in $\mathcal{X}$, with $|\mathcal{X}| = K$.

(5) Maximum a Posterior (MAP) Estimation: $\hat{x}_{0:T} = \arg\max_{x_{0:T}} p(x_{0:T} \mid y_{1:T}) = \arg\max_{x_{0:T}} p(x_{0:T}, y_{1:T})$.

Algorithm 8: Viterbi's Backward-Forward Recursion

```
Initialize the message passed from the end of chain: $m_T(i) = 1$ for
 $i \in \mathcal{X} = \{1, \cdots, K\}$.
for $t = T, \cdots, 1$ do
    for $i \in \mathcal{X}$ do
        $m_{t-1}(i) = \max_{x_{t:T}} \left\{ \prod_{k=t}^{T} p\left(y_k \mid x_k\right) p\left(x_k \mid x_{k-1}\right) \right\} \quad = \max_{x_t} p\left(y_t \mid x_t\right) p\left(x_t \mid x_{t-1}\right) m_t\left(x_t\right)$
        $= \max_{j=1,\ldots,K} g_j\left(y_t\right) A_{i,j} m_t(j)$
    end
end
Set $\hat{x}_0 = \arg\max_{x_0} \left(p\left(x_0\right) m_0\left(x_0\right)\right) = \underset{i=1\ldots,K}{\arg\max} \{\mu(i) m_0(i)\}$
for $t = 1, \cdots, T$ do
        $\hat{x}_t = \underset{x_{2:T}}{\arg\max} p\left(\hat{x}_0, x_{1:T}, y_{1:T}\right)$
        $= \underset{x_{2:T}}{\arg\max} p\left(\hat{x}_0\right) p\left(y_1 \mid x_1\right) p\left(x_1 \mid \hat{x}_0\right) \prod_{t=2}^{T} p\left(y_t \mid x_t\right) p\left(x_t \mid x_{t-1}\right)$
        $= \underset{i=1\ldots,K}{\arg\max} g_i\left(y_t\right) A_{\hat{x}_{t-1},i} m_t(i)$
end
```

4.1.4 Estimation of Transition Probability

(1) Fully Observed Case: when both the data $y_{1:T}$ and the hidden states $x_{0:T}$ are known, the MLE of $A_{i,j} = \mathbb{P}(X_t = j | X_{t-1} = i)$:

$$\widehat{A}_{i,j} = \frac{n_{i,j}}{\sum_{\ell \in \mathcal{X}} n_{i,\ell}}$$

with $n_{i,j} = \sum_{t=1}^{T} I\left(x_t = j, x_{t-1} = i\right)$.

(2) Unsupervised Case: when both the data $y_{1:T}$ known but the hidden states $x_{0:T}$ unknown, want to find $\widehat{A} = \underset{A}{\arg\max} \log p\left(y_{1:T}; A\right)$ via Algorithm 9.

```
Algorithm 9: Baum-Welch (a special case of Expectation-Maximization algorithm)
for $k = 1, \cdots, K$ do
    E-Step: Compute
        $Q\left(A; A^{(k-1)}\right) = \mathbb{E}\left[\log p\left(X_{0:T}, y_{1:T}; A\right) \mid y_{1:T}, A^{(k-1)}\right]$, with $p\left(x_{0:T}, y_{1:T}; A\right) = \mu_{x_0} \prod_{i,j \in \mathcal{X}} A_{i,j}^{n_{i,j}} \prod_{t=1}^{T} g_{y_t}(x_t)$
        $= \mathbb{E}\left[\log \mu_{x_0} + \sum_{i,j \in \mathcal{X}} n_{i,j} \log A_{i,j} + \sum_{t=1}^{T} \log g_{y_t}\left(x_t\right) \mid y_{1:T}, A^{(k-1)}\right]$
        $= \sum_{i,j \in \mathcal{X}} \mathbb{E}\left[N_{i,j} \mid y_{1:T}, A^{(k-1)}\right] \log A_{i,j} + C$
        $= \sum_{t=1}^{T} \mathbb{P}\left(X_t = j, X_{t-1} = i \mid y_{1:T}; A^{(k-1)}\right) + C \quad \because N_{i,j} = \sum_{t=1}^{T} I\left(X_t = j, X_{t-1} = i\right)$
    • where $p\left(x_t, x_{t-1} \mid y_{1:T}; A^{(t-1)}\right) \propto \alpha_{t-1}\left(x_{t-1}\right) p\left(y_t \mid x_t\right) p\left(x_t \mid x_{t-1}\right) \beta_t\left(x_t\right)$ can be
      computed via backward-forward recursion (Alg 8).
    M-Step: Compute
        $A_{i,j}^{(k)} = \arg\max_{A} Q(A; A^{(k-1)})$
        $= \arg\max_{A_{i,j}} \left\{\mathbb{E}\left[N_{i,j} \mid y_{1:T}, A^{(k-1)}\right] \log A_{i,j}\right\} = \frac{\mathbb{E}\left[N_{i,j} \mid y_{1:T}, A^{(k-1)}\right]}{\sum_{\ell} \mathbb{E}\left[N_{i,\ell} \mid y_{1:T}, A^{(k-1)}\right]}$
end
```

.

4.2 Continuous-state Hidden Markov models

• When hidden states are continuous, e.g. Gaussian state-space model.

4.2.1 Linear Gaussian System

Multi-Variate Gaussian:

$$\mathcal{N}(x; \mu, \Sigma) := \frac{1}{(2\pi)^{d_x/2}\sqrt{|\Sigma|}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}$$

Conditional Gaussian:

$$\begin{aligned} X \mid Y = y &\sim \mathcal{N}\left(\mu_{x|y}, \Sigma_{x|y}\right) \\ \Sigma_{x|y} &= \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} \\ \mu_{x|y} &= \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}\left(y - \mu_y\right) \end{aligned}$$

where, $(X, Y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ is jointly Gaussian with:
- *marginals*: $X \sim N(\mu_x, \Sigma_{xx})$ and $Y \sim N(\mu_y, \Sigma_{yy})$;

- *mean and covariance*: $\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}$

Prop 4.2:
Suppose:
(i) $X \in \mathbb{R}^{d_x}$ s.t. $X \sim N(\mu_x, \Sigma_{xx})$;
(ii) $W \in \mathbb{R}^{d_w}$ s.t. $W \sim N(0, R)$ and $X \perp W$;
(iii) $Y \in \mathbb{R}^{d_y}$ s.t. $Y = HX + b + W$, with H a $d_y \times d_x$ matrix, b a $d_y \times 1$ vector.
Then,

$$Y \sim \mathcal{N}(\mu_y, \Sigma_{yy}) \quad , X \mid Y = y \sim \mathcal{N}\left(\mu_{x|y}, \Sigma_{x|y}\right)$$

with,

$$\begin{aligned} \mu_y &= H\mu_x + b \\ \Sigma_{yy} &= H\Sigma_{xx}H^\top + R \\ \Sigma_{x|y} &= \Sigma_{xx} - \Sigma_{xx}H^\top \left(H\Sigma_{xx}H^\top + R\right)^{-1} H\Sigma_{xx} \\ \mu_{x|y} &= \mu_x + \Sigma_{xx}H^\top \left(H\Sigma_{xx}H^\top + R\right)^{-1} (y - H\mu_x - b) \end{aligned}$$

Proof: We know that $Y \sim N(\mu_y, \Sigma_{yy})$, with

$$\begin{aligned} \mu_y &= \mathbb{E}[Y] = H\mathbb{E}[X] + b + \mathbb{E}[W] = H\mu_x + b \\ \Sigma_{yy} &= \operatorname{cov}[Y] = H(\operatorname{cov}[X])H^\top + \operatorname{cov}[W] = H\Sigma_{xx}H^\top + R \end{aligned}$$

Covariance between X and Y is

$$\begin{aligned} \operatorname{cov}[X, Y] &= \mathbb{E}\left[(X - \mu_x)(Y - \mu_y)\right]^\top = \mathbb{E}\left[(X - \mu_x)(H(X - \mu_x) + W)\right]^\top \\ &= \operatorname{cov}[X]H^\top = \Sigma_{xx}H^\top \quad \because X \perp W \end{aligned}$$

Therefore (X, Y) is jointly Gaussian with: $\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xx}H^\top \\ H\Sigma_{xx} & \Sigma_{yy} \end{pmatrix}$.
By Conditional Gaussian, $X|Y \sim N(\mu_{x|y}, \Sigma_{x|y})$ with:

$$\begin{aligned} \Sigma_{x|y} &= \Sigma_{xx} - \Sigma_{xx}H^\top\Sigma_{yy}^{-1}H\Sigma_{xx} \\ &= \Sigma_{xx} - \Sigma_{xx}H^\top \left(H\Sigma_{xx}H^\top + R\right)^{-1} H\Sigma_{xx} \\ \mu_{x|y} &= \mu_x + \Sigma_{xx}H^\top\Sigma_{yy}^{-1}(y - \mu_y) \\ &= \mu_x + \Sigma_{xx}H^\top \left(H\Sigma_{xx}H^\top + R\right)^{-1} (y - H\mu_x - b) \quad \text{[EOP]} \end{aligned}$$

4.2.2 Dynamic Linear Gaussian state-space models (SSM)

State Model: $X_t = F_tX_{t-1} + G_tV_t$
Observation Model: $Y_t = H_tX_t + W_t$
where,
- X_t is the hidden state at time t, with $X_0 \sim N(\mu_0, \Sigma_0)$;
- Y_t is the observation at time t;
- $V_t \sim N(0, Q_t)$ is the state noise at time t;

- $W_t \sim N(0, R_t)$ is the observation noise at time t;
- $(X_0, V_{1:T}, W_{1:T})$ are independent;
- F_t is the $d_x \times d_x$ state transition matrix;
- G_t is the $d_x \times d_v$ noise transfer matrix;
- H_t is the $d_y \times d_x$ observation matrix.

Joint pdf: If $G_t Q_t G_t^T$ factorises,

$$p\left(x_{0:T}, y_{1:T}\right) = p\left(x_0\right) \prod_{t=1}^{T} p\left(y_t \mid x_t\right) p\left(x_t \mid x_{t-1}\right)$$

where $p\left(x_t \mid x_{t-1}\right) = \mathcal{N}\left(x_t; F_t x_{t-1}, G_t Q_t G_t^{\top}\right), \quad p\left(y_t \mid x_t\right) = \mathcal{N}\left(y_t; H_t x_t, R_t\right)$
(Examples omitted, see P15 on the notes).

4.2.3 Inference in dynamic linear Gaussian SSMs

(1) The Kalman filter: want to compute $p(x_t|y_{1:t})$, via prediction-update recursion.

$$p\left(x_t \mid y_{1:t-1}\right) = \mathcal{N}\left(x_t; \mu_{t|t-1}, \Sigma_{t|t-1}\right) \text{ (prediction)}$$
$$p\left(x_t \mid y_{1:t}\right) = \mathcal{N}\left(x_t; \mu_{t|t}, \Sigma_{t|t}\right) \text{ (update)}$$

where $(\mu_{t|t-1}, \Sigma_{t|t-1})$ and $\mu_{t|t}, \Sigma_{t|t}$ follow the recursion below (Alg 10)

Algorithm 10: Kalman Prediction-Update recursion

```
Initialize $(\mu_0, \Sigma_0)$ as the starting stage.
for $t = 1, \cdots, T$ do
    Prediction step:
      $\mu_{t|t-1} := \mathbb{E}\left[X_t \mid Y_{1:t-1} = y_{1:t-1}\right] = F_t \mu_{t-1|t-1}$
      $\Sigma_{t|t-1} := \mathbb{E}\left[\left(X_t - \mu_{t|t-1}\right)\left(X_t - \mu_{t|t-1}\right)^{\top} \mid Y_{1:t-1} = y_{1:t-1}\right] = F_t \Sigma_{t-1|t-1} F_t^{\top} + G_t Q_t G_t^{\top}$
    Update/correction step:
      $\mu_{t|t} := \mathbb{E}\left[X_t \mid Y_{1:t} = y_{1:t}\right] = \mu_{t|t-1} + K_t \nu_t$
      $\Sigma_{t|t} := \mathbb{E}\left[\left(X_t - \mu_{t|t}\right)\left(X_t - \mu_{t|t}\right)^{\top} \mid Y_{1:t} = y_{1:t}\right] = \left(I - K_t H_t\right) \Sigma_{t|t-1}$
    where:
      (i) Residual/innovation: $\nu_t := y_t - \hat{y}_{t|t-1}$, with $\hat{y}_{t|t-1} := \mathbb{E}\left[Y_t \mid Y_{1:t-1} = y_{1:t-1}\right] = H_t \mu_{t|t-1}$
      (ii) Kalman gain$K_t = \Sigma_{t|t-1} H_t^{\top} S_t^{-1}$,
      with $S_t := \mathbb{E}\left[\left(Y_t - \hat{y}_{t|t-1}\right)\left(Y_t - \hat{y}_{t|t-1}\right)^{\top} \mid Y_{1:t-1} = y_{1:t-1}\right] = H_t \Sigma_{t|t-1} H_t^{\top} + R_t$
end
```

Proof: omitted, see P16 on the notes.

(2) Kalman Smoother: want to compute $p(x_t|y_{1:T})$ via backward recursion.

$$p\left(x_t \mid y_{1:T}\right) = \mathcal{N}\left(x_t; \mu_{t|T}, \Sigma_{t|T}\right)$$

Algorithm 11: Kalman Backward Recursion

```
Initialize $(\mu_{T|T}, \Sigma_{T|T})$ as the starting stage and go backward.
for $t = T-1, \cdots, 0$ do
    $\mu_{t|T} = \mu_{t|t} + J_t \left(\mu_{t+1|T} - \mu_{t+1|t}\right)$
    $\Sigma_{t|T} = \Sigma_{t|t} + J_t \left(\Sigma_{t+1|T} - \Sigma_{t+1|t}\right) J_t^\top$
    with the backward Kalman gain: $J_t = \Sigma_{t|t} F_{t+1}^\top \Sigma_{t+1|t}^{-1}$
end
```