Controlling for confounders through approximate sufficiency

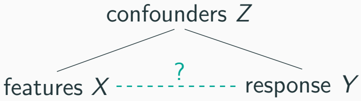
Rina Foygel Barber (joint with Lucas Janson)

http://www.stat.uchicago.edu/~rina/

Collaborator

Lucas Janson (Harvard U.)

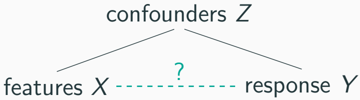
Intro: testing conditional independence



Classical (parametric) approach:

- Assume a parametric model such as $Y \mid X, Z \sim f(\cdot\,; \alpha^\top X + \beta^\top Z)$
- Parametric inference to test $H_0 : \alpha = 0$

Intro: testing conditional independence



Classical (parametric) approach:

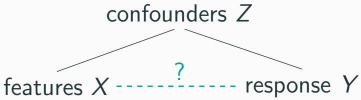
- Assume a parametric model such as $Y \mid X, Z \sim f(\cdot\,; \alpha^\top X + \beta^\top Z)$
- Parametric inference to test $H_0 : \alpha = 0$

Model-X approach a.k.a. Conditional Randomization Test (Candès et al 2018)

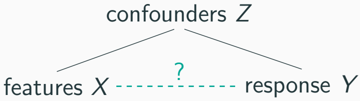
- Known distribution of $X \mid Z$ (distrib. of Y unknown)
- Choose function $T(X; Y, Z)$ that measures association
- Resample copies $\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)} \stackrel{\text{iid}}{\sim}$ (distrib. of $X \mid Z$)

$$\rightsquigarrow \quad \text{pval} = \frac{1 + \sum_m \mathbb{1}\{T(\tilde{X}^{(m)}; Y, Z) \geq T(X; Y, Z)\}}{1 + M}$$

Intro: testing conditional independence

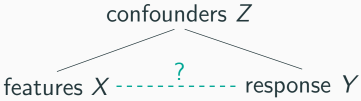


Intro: testing conditional independence



Model-X approach via sufficient statistics (Huang & Janson 2019)

- Distribution of $X \mid Z$ is only partially known
- By conditioning on sufficient statistic $S(X, Z)$,
  can resample copies $\tilde{X}^{(1)}, \dots, \tilde{X}^{(M)} \overset{\text{iid}}{\sim}$ (distrib. of $X \mid S(X, Z)$)
  & compute p-value for test statistic T as before



Model-X approach via sufficient statistics (Huang & Janson 2019)

- Distribution of $X \mid Z$ is only partially known
- By conditioning on sufficient statistic $S(X, Z)$,
  can resample copies $\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)} \stackrel{\text{iid}}{\sim}$ (distrib. of $X \mid S(X, Z)$)
  & compute p-value for test statistic T as before
- Example: canonical GLMs
  — $X_i \sim \exp\left\{X_i \cdot Z_i^\top \theta - a(Z_i^\top \theta)\right\}$, $i = 1, \ldots, n$, with θ unknown
  — $S(X, Z) = \sum_i X_i Z_i$ is suff. stat. for $X = (X_1, \ldots, X_n)$

Intro: testing goodness-of-fit (GoF)

More generally...

Goodness-of-fit test

Testing H_0: $X \sim P_\theta$ for some $\theta \in \Theta$,

where $\{P_\theta : \theta \in \Theta\}$ is a parametric family

Intro: testing goodness-of-fit (GoF)

More generally...

Goodness-of-fit test

Testing H_0: $X \sim P_\theta$ for some $\theta \in \Theta$,
where $\{P_\theta : \theta \in \Theta\}$ is a parametric family

Conditional independence testing can be a special case:

- Assume $X \mid Z \sim P_\theta(\cdot|Z)$ for some $\theta \in \Theta$
- Null hypothesis $H_0 : X \perp\!\!\!\perp Y \mid Z$
- Equivalently... H_0: $X \mid Y, Z \sim P_\theta(\cdot|Z)$ for some $\theta \in \Theta$
- Note: we condition on Y and Z (i.e., treat as fixed)

Intro: testing goodness-of-fit (GoF)

A general framework:

- Choose any test statistic $T : \mathcal{X} \to \mathbb{R}$
- Draw copies $\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}$
- Compute rank-based p-value

$$\text{pval} = \frac{1 + \sum_m \mathbb{1}\{T(\tilde{X}^{(m)}) \geq T(X)\}}{1 + M}$$

- If $X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}$ are exchangeable under H_0 $\rightsquigarrow$ p-value is valid

Co-sufficient sampling (CSS)

Co-sufficient sampling

Sample copies $\tilde{X}^{(m)} \sim$ (distrib. of $X \mid S(X)$),
where $S(X)$ is a sufficient statistic for the family $\{P_\theta : \theta \in \Theta\}$

Can be applied to:

1. Test goodness-of-fit (GoF)
   (Engen & Lillegård 1997, Lockhart et al 2007, Stephens 2012, Hazra 2013 ....)
2. Test conditional independence (special case of GoF)
   (Rosenbaum 1984, Kolassa 2003, Huang & Janson 2019)
3. Construct conf. intervals for a parameter of interest
   (by inverting GoF tests)

Co-sufficient sampling (CSS)

Co-sufficient sampling

Sample copies $\tilde{X}^{(m)} \sim$ (distrib. of $X \mid S(X)$),

where $S(X)$ is a sufficient statistic for the family $\{P_\theta : \theta \in \Theta\}$

Co-sufficient sampling (CSS)

Co-sufficient sampling

Sample copies $\tilde{X}^{(m)} \sim$ (distrib. of $X \mid S(X)$),

where $S(X)$ is a sufficient statistic for the family $\{P_\theta : \theta \in \Theta\}$

Permutation tests are an example of CSS

- H_0: $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} \mathcal{D}$ for $\mathcal{D} \in$ (some set)
- The order statistics $X_{(1)} \leq \cdots \leq X_{(n)}$ are sufficient under the null
- Permutation test $\Leftrightarrow$ resampling X conditional on order statistics
- Application: testing $X \perp\!\!\!\perp Y$

  H_0: conditional on $Y_1, \ldots, Y_n$, it holds that $X_1, \ldots, X_n$ are i.i.d.

Co-sufficient sampling (CSS)

Limitation of co-sufficient sampling... no power in many settings!

Example—logistic model:

- $X = (X_1, \ldots, X_n) \in \{0,1\}^n$, $Z = (Z_1, \ldots, Z_n) \in (\mathbb{R}^k)^n$
- If the Z_i's are in general position,

  then $\sum_i X_i Z_i \in \mathbb{R}^k$ uniquely determines X

  (so if we resample, will have $\tilde{X}^{(1)} = \cdots = \tilde{X}^{(M)} = X \rightsquigarrow$ zero power)

Co-sufficient sampling (CSS)

Limitation of co-sufficient sampling... no power in many settings!

Co-sufficient sampling (CSS)

Limitation of co-sufficient sampling... no power in many settings!

For many other models, the minimal sufficient statistic $S(X)$ is essentially the data itself, e.g.,

- Mixture of Gaussians or mixture of GLMs
- Non-canonical GLMs
- Heavy tailed distributions (e.g., multivariate t)
- Models with missing or corrupted data

Approximate sufficiency

For a family $\{P_\theta : \theta \in \Theta\}$, a function $S(X)$ is a *sufficient statistic* if

$$(\text{distrib. of } X \mid S(X),\ X \sim P_\theta) = (\text{distrib. of } X \mid S(X),\ X \sim P_{\theta'}) \quad \forall \theta, \theta'.$$

Asymptotic sufficiency: (Le Cam, Wald, ...)

Informally...

$$(\text{distrib. of } X \mid S(X),\ X \sim P_\theta) \approx (\text{distrib. of } X \mid S(X),\ X \sim P_{\theta'}) \quad \forall \theta, \theta'.$$

- Under regularity conditions, $S(X) = \widehat{\theta}_{\mathsf{MLE}}(X)$ is asymp. suff.

Approximate co-sufficient sampling (aCSS)

Main idea:

- Let $\widehat{\theta} \in \Theta$ be an approximate MLE given the data X
- Let $p_\theta(\cdot|\widehat{\theta})$ = distrib. of $X \mid \widehat{\theta}$, if marginally $X \sim P_\theta$
  $\rightsquigarrow$ under the null, $X \mid \widehat{\theta} \sim p_{\theta_0}(\cdot|\widehat{\theta})$ for the unknown true θ_0
- Sample copies $\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}$ from $\underbrace{p_{\widehat{\theta}}(\cdot|\widehat{\theta}) \approx p_{\theta_0}(\cdot|\widehat{\theta})}_{\text{by approx. sufficiency}}$

  $X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)} \approx$ exchangeable under H_0 $\rightsquigarrow$ p-value is $\approx$ valid

Approximate co-sufficient sampling (aCSS)

Distance to exchangeability

$$\mathrm{d}_{\mathrm{exch}}(X, \tilde{X}^{(1)}, \dots, \tilde{X}^{(M)}) = \inf_{\substack{\text{Exch. distrib.}\\ \mathcal{D} \text{ on } \mathcal{X}^{M+1}}} \left\{ \mathrm{d}_{\mathrm{TV}}\Big((X, \tilde{X}^{(1)}, \dots, \tilde{X}^{(M)}), \mathcal{D} \Big) \right\}$$

For any test statistic $T(X)$, the p-value

$$\text{pval} = \frac{1 + \sum_m \mathbb{1}\{T(\tilde{X}^{(m)}) \geq T(X)\}}{1 + M}$$

satisfies

$$\mathbb{P}\left\{\text{pval} \leq \alpha\right\} \leq \alpha + \mathrm{d}_{\mathrm{exch}}(X, \tilde{X}^{(1)}, \dots, \tilde{X}^{(M)}).$$

aCSS algorithm

- Step 1: choose a test statistic $T : \mathcal{X} \to \mathbb{R}$
- Step 2: observe data X, and compute an approximate MLE $\widehat{\theta}$
- Step 3: sample copies $\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}$ from $\approx$ distribution of $X \mid \widehat{\theta}$
- Step 4: compute a rank-based p-value to test H_0:

$$\text{pval} = \frac{1 + \sum_m \mathbb{1}\{T(\tilde{X}^{(m)}) \geq T(X)\}}{1 + M}$$

aCSS algorithm

- Step 1: choose a test statistic $T : \mathcal{X} \to \mathbb{R}$
- Step 2: observe data X, and compute an approximate MLE $\widehat{\theta}$
- Step 3: sample copies $\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}$ from $\approx$ distribution of $X \mid \widehat{\theta}$
- Step 4: compute a rank-based p-value to test H_0:

$$\mathrm{pval} = \frac{1 + \sum_m \mathbb{1}\{T(\tilde{X}^{(m)}) \geq T(X)\}}{1 + M}$$

aCSS algorithm

- Step 2: observe data X, and compute an approximate MLE $\widehat{\theta}$

Ideally would like to minimize

$$\mathcal{L}(\theta; X, W) = \underbrace{\mathcal{L}(\theta; X)}_{\substack{\text{penalized neg. log-likelihood} \\ -\log f(X;\theta)+\mathcal{R}(\theta)}} + \underbrace{\sigma \cdot W^\top \theta}_{\substack{\text{perturb with } W \sim \mathcal{N}(0, \frac{1}{d}\mathrm{I}_d) \\ (\text{choose } \sigma \ll n^{1/2})}}$$

(see also Tian & Taylor 2018—random perturbation for selective inference)

---

aCSS algorithm

- Step 2: observe data X, and compute an approximate MLE $\widehat{\theta}$

Ideally would like to minimize

$$\mathcal{L}(\theta; X, W) = \underbrace{\mathcal{L}(\theta; X)}_{\substack{\text{penalized neg. log-likelihood} \\ -\log f(X;\theta)+\mathcal{R}(\theta)}} + \underbrace{\sigma \cdot W^\top \theta}_{\substack{\text{perturb with } W \sim \mathcal{N}(0, \frac{1}{d}\mathrm{I}_d) \\ (\text{choose } \sigma \ll n^{1/2})}}$$

(see also Tian & Taylor 2018—random perturbation for selective inference)

But... what if nonconvex? what if no global minimum?

— Function $\widehat{\theta} : \mathcal{X} \times \mathbb{R}^d \to \Theta$, returns $\widehat{\theta}(X, W)$.

— If $\widehat{\theta}(X, W)$ is a strict SOSP of $\mathcal{L}(\theta; X, W)$, proceed to next step.

— Otherwise return $\tilde{X}^{(1)} = \cdots = \tilde{X}^{(M)} = X \rightsquigarrow$ pval $= 1$.

aCSS algorithm

- Step 3: sample copies $\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}$ from $\approx$ distribution of $X \mid \widehat{\theta}$

aCSS algorithm

- Step 3: sample copies $\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}$ from $\approx$ distribution of $X \mid \widehat{\theta}$

Density of $X \mid \widehat{\theta}$, conditional on the event that $\widehat{\theta}(X, W)$ is strict SOSP:

$$\propto f(x; \theta_0) \cdot \exp\left\{-\frac{\|\nabla_\theta \mathcal{L}(\widehat{\theta}; x)\|}{2\sigma^2/d}\right\} \cdot \det\left(\nabla^2_\theta \mathcal{L}(\widehat{\theta}; x)\right) \cdot \mathbb{1}_{x \in \mathcal{X}_{\widehat{\theta}}}$$

support of $X|\widehat{\theta}$

aCSS algorithm

- Step 3: sample copies $\tilde{X}^{(1)}, \dots, \tilde{X}^{(M)}$ from $\approx$ distribution of $X \mid \widehat{\theta}$

Density of $X \mid \widehat{\theta}$, conditional on the event that $\widehat{\theta}(X, W)$ is strict SOSP:

$$\propto f(x; \theta_0) \cdot \exp\left\{ -\frac{\|\nabla_\theta \mathcal{L}(\widehat{\theta}; x)\|}{2\sigma^2/d} \right\} \cdot \det\left(\nabla^2_\theta \mathcal{L}(\widehat{\theta}; x)\right) \cdot \mathbb{1}_{x \in \mathcal{X}_{\widehat{\theta}}}$$

support of $X|\widehat{\theta}$

θ_0 unknown $\rightsquigarrow$ use $\widehat{\theta}$ as plug-in estimate:

$$\propto f(x; \widehat{\theta}) \cdot \exp\left\{ -\frac{\|\nabla_\theta \mathcal{L}(\widehat{\theta}; x)\|}{2\sigma^2/d} \right\} \cdot \det\left(\nabla^2_\theta \mathcal{L}(\widehat{\theta}; x)\right) \cdot \mathbb{1}_{x \in \mathcal{X}_{\widehat{\theta}}}$$

aCSS algorithm

- Step 3: sample copies $\tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}$ from $\approx$ distribution of $X \mid \widehat{\theta}$

Density of $X \mid \widehat{\theta}$, conditional on the event that $\widehat{\theta}(X, W)$ is strict SOSP:

$$\propto f(x; \theta_0) \cdot \exp\left\{-\frac{\|\nabla_\theta \mathcal{L}(\widehat{\theta}; x)\|}{2\sigma^2/d}\right\} \cdot \det\left(\nabla^2_\theta \mathcal{L}(\widehat{\theta}; x)\right) \cdot \mathbb{1}_{x \in \mathcal{X}_{\widehat{\theta}}}$$

support of $X|\widehat{\theta}$

θ_0 unknown $\rightsquigarrow$ use $\widehat{\theta}$ as plug-in estimate:

$$\propto f(x; \widehat{\theta}) \cdot \exp\left\{-\frac{\|\nabla_\theta \mathcal{L}(\widehat{\theta}; x)\|}{2\sigma^2/d}\right\} \cdot \det\left(\nabla^2_\theta \mathcal{L}(\widehat{\theta}; x)\right) \cdot \mathbb{1}_{x \in \mathcal{X}_{\widehat{\theta}}}$$

If sampling directly is impossible,

can use an exchangeable form of MCMC (Besag & Clifford 1989)

Type I error guarantee

Assumption 1: regularity conditions

- $\Theta \subseteq \mathbb{R}^d$ convex & open
- P_θ has positive density $f(\cdot;\theta)$ w.r.t. base measure $\nu_{\mathcal{X}}$ for all $\theta \in \Theta$
- Log-likelihood $\log f(x;\theta)$ & penalty $\mathcal{R}(\theta)$ are continuously twice diff.

Type I error guarantee

Assumption 2: approximate MLE

For $X \sim P_{\theta_0}$ and $W \sim \mathcal{N}(0, \frac{1}{d}\mathrm{I}_d)$, with prob. at least $1-\delta$,

$$\|\widehat{\theta}(X,W) - \theta_0\| \leq r \quad \text{and} \quad \widehat{\theta}(X,W) \text{ is a strict SOSP of } \mathcal{L}(\theta; X, W).$$

Assumption 3: Hessian of the log-likelihood

$$\mathbb{E}\left[\exp\left\{\sup_{\theta \in \mathbb{B}(\theta_0, r)\cap\Theta} r^2 \|\nabla^2 \log f(X;\theta) - \mathbb{E}\left[\nabla^2 \log f(X;\theta)\right]\|\right\}\right] \leq e^{\varepsilon}$$

Type I error guarantee

Assumption 2: approximate MLE

For $X \sim P_{\theta_0}$ and $W \sim \mathcal{N}(0, \frac{1}{d}\mathrm{I}_d)$, with prob. at least $1-\delta$,

$$\|\widehat{\theta}(X,W) - \theta_0\| \leq r \quad \text{and} \quad \widehat{\theta}(X,W) \text{ is a strict SOSP of } \mathcal{L}(\theta; X, W).$$

Assumption 3: Hessian of the log-likelihood

$$\mathbb{E}\left[\exp\left\{\sup_{\theta\in\mathbb{B}(\theta_0,r)\cap\Theta} r^2\|\nabla^2 \log f(X;\theta) - \mathbb{E}\left[\nabla^2 \log f(X;\theta)\right]\|\right\}\right] \leq e^{\varepsilon}$$

In standard settings with n independent observations...

$$r, \varepsilon, \delta = \widetilde{\mathcal{O}}(n^{-1/2})$$

Type I error guarantee

Theorem

Under Assumptions 1, 2, & 3, the copies produced by aCSS satisfy

$$\mathsf{d}_{\mathsf{exch}}(X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}) \leq 3\sigma r + \delta + \varepsilon$$

under H_0.

Therefore, for any test statistic T, Type I error for testing H_0 satisfies

$$\mathbb{P}\left\{\mathsf{pval} \leq \alpha\right\} \leq \alpha + 3\sigma r + \delta + \varepsilon$$

Type I error guarantee

Theorem

Under Assumptions 1, 2, & 3, the copies produced by aCSS satisfy

$$\mathsf{d}_{\mathsf{exch}}(X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}) \leq 3\sigma r + \delta + \varepsilon$$

under H_0.

Therefore, for any test statistic T, Type I error for testing H_0 satisfies

$$\mathbb{P}\left\{\mathsf{pval} \leq \alpha\right\} \leq \alpha + 3\sigma r + \delta + \varepsilon$$

Excess Type I error should be $o(1)$...

- $r, \delta, \varepsilon \asymp n^{-1/2}$ from the assumptions
- $\sigma =$ noise level, chosen by analyst
  $\rightarrow$ choose $\sigma \asymp n^c$ for some $c \in [0, \frac{1}{2})$

Examples

Examples where CSS has no power, but aCSS assumptions hold:

- Canonical GLMs such as logistic regression (low-dim.):

$$X_i \overset{\perp\!\!\!\perp}{\sim} \text{Bernoulli}\left(\frac{e^{Z_i^\top \beta}}{1+e^{Z_i^\top \beta}}\right) \text{ for unknown } \beta$$

- Two-sample difference-of-means (the Behrens–Fisher problem):

$$X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_X, \sigma_X^2), \quad Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2), \quad \text{test } H_0 : \mu_X = \mu_Y$$

(An aCSS-like approach for this problem was considered by Lillegård 2001)

Examples

Examples where CSS has no power, but aCSS assumptions hold:

- Spatial process on integer lattice: for unknown ρ,

  $X \sim \mathcal{N}(0, \Sigma)$ where $\Sigma_{ij} = \rho^{D_{ij}}$ for known pairwise distances D_{ij}

- Multivariate t distribution (low-dim.):

  $$X_i \stackrel{\text{iid}}{\sim} t_\gamma(0, \Sigma) \text{ for known } \gamma \text{ \& unknown } \Sigma$$

- And maybe missing data, latent variables, and more ...

Simulations

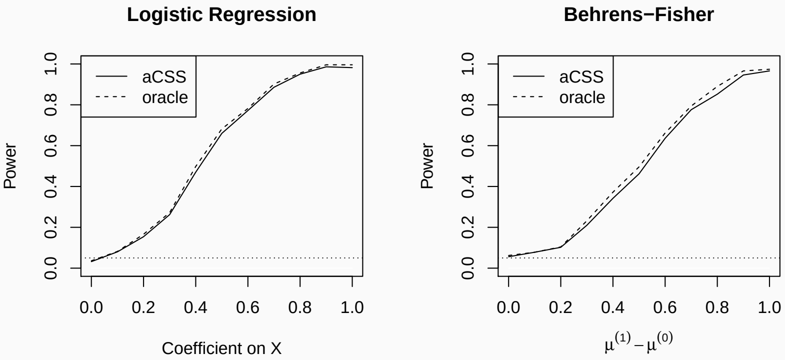
Compare to oracle method that knows θ_0:

- Sample copies $\tilde{X}^{(m)} \overset{\text{iid}}{\sim} P_{\theta_0}$
- Compute p-value with same statistic $T(x)$

Simulations

Compare to oracle method that knows θ_0:

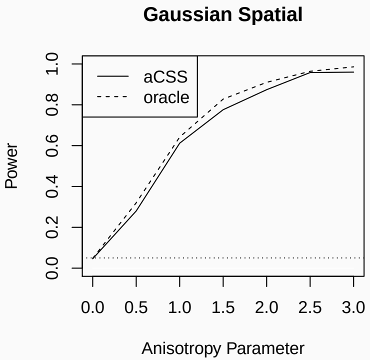
- Sample copies $\tilde{X}^{(m)} \overset{\text{iid}}{\sim} P_{\theta_0}$
- Compute p-value with same statistic $T(x)$

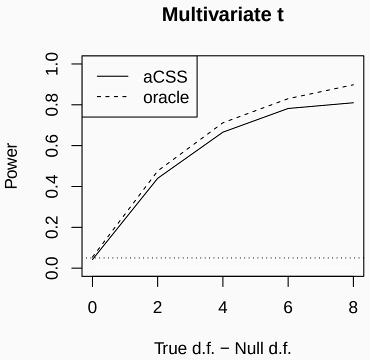


Simulations

Compare to oracle method that knows θ_0:

- Sample copies $\tilde{X}^{(m)} \overset{\text{iid}}{\sim} P_{\theta_0}$
- Compute p-value with same statistic $T(x)$





Sampling

Recall: need to sample copies $\tilde{X}^{(m)}$ from

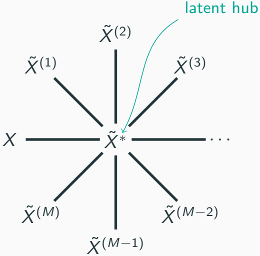
$$\propto f(x;\widehat{\theta}) \cdot \exp\left\{-\frac{\|\nabla_\theta \mathcal{L}(\widehat{\theta};x)\|}{2\sigma^2/d}\right\} \cdot \det\left(\nabla^2_\theta \mathcal{L}(\widehat{\theta};x)\right) \cdot \mathbb{1}_{x\in\mathcal{X}_{\widehat{\theta}}}$$

Sampling

Recall: need to sample copies $\tilde{X}^{(m)}$ from

$$\propto f(x;\widehat{\theta}) \cdot \exp\left\{-\frac{\|\nabla_\theta \mathcal{L}(\widehat{\theta};x)\|}{2\sigma^2/d}\right\} \cdot \det\left(\nabla^2_\theta \mathcal{L}(\widehat{\theta};x)\right) \cdot \mathbb{1}_{x\in\mathcal{X}_{\widehat{\theta}}}$$

Two exchangeable MCMC strategies (Besag & Clifford 1989)



$$\underbrace{\tilde{X}^{(4)} — \tilde{X}^{(2)} — X — \tilde{X}^{(1)} — \cdots — \tilde{X}^{(M)} — \tilde{X}^{(3)}}_{\text{Random permutation of } M+1 \text{ positions}}$$

- Run Metropolis–Hastings, where $f(x;\widehat{\theta})$ stationary for proposal distrib.
- e.g., if X consists of n indep. observations (i.e., $f(x;\widehat{\theta}) = \prod_{i=1}^n f_i(x_i;\widehat{\theta})$), can choose proposal distrib. $=$ resample s of n observations

Proof sketch for Theorem

Need to bound $\mathrm{d}_{\mathrm{exch}}(X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)})$

---

(1) Calculate joint distribution:

$$\begin{cases} \widehat{\theta} & \sim (\text{marginal distrib. of } \widehat{\theta}) \\ X \mid \widehat{\theta} & \sim p_{\theta_0}(\cdot|\widehat{\theta}) \\ \tilde{X}^{(m)} \mid X, \widehat{\theta} & \sim p_{\widehat{\theta}}(\cdot|\widehat{\theta}) \end{cases}$$

$$\Longrightarrow \quad \mathrm{d}_{\mathrm{exch}}(X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}) \;\leq\; \mathbb{E}_{\widehat{\theta}}\left[\mathrm{d}_{\mathrm{TV}}\big(p_{\theta_0}(\cdot|\widehat{\theta}), p_{\widehat{\theta}}(\cdot|\widehat{\theta})\big)\right]$$

Proof sketch for Theorem

(2) To bound d_{TV}:

$$\frac{p_{\widehat{\theta}}(X|\widehat{\theta})}{p_{\theta_0}(X|\widehat{\theta})} \propto \frac{f(X;\widehat{\theta}\,)}{f(X;\theta_0)} \quad \Rightarrow \quad \frac{p_{\widehat{\theta}}(X|\widehat{\theta})}{p_{\theta_0}(X|\widehat{\theta})} = \frac{\frac{f(X;\widehat{\theta}\,)}{f(X;\theta_0)}}{\mathbb{E}_{p_{\theta_0}(\cdot|\widehat{\theta})}\left[\frac{f(X;\widehat{\theta}\,)}{f(X;\theta_0)}\right]}$$

Proof sketch for Theorem

(2) To bound d_{TV}:

$$\frac{p_{\widehat{\theta}}(X|\widehat{\theta})}{p_{\theta_0}(X|\widehat{\theta})} \propto \frac{f(X;\widehat{\theta})}{f(X;\theta_0)} \quad \Rightarrow \quad \frac{p_{\widehat{\theta}}(X|\widehat{\theta})}{p_{\theta_0}(X|\widehat{\theta})} = \frac{\frac{f(X;\widehat{\theta})}{f(X;\theta_0)}}{\mathbb{E}_{p_{\theta_0}(\cdot|\widehat{\theta})}\left[\frac{f(X;\widehat{\theta})}{f(X;\theta_0)}\right]}$$

$$\Rightarrow \quad \mathrm{d}_{\mathrm{TV}}\left(p_{\theta_0}(\cdot|\widehat{\theta}), p_{\widehat{\theta}}(\cdot|\widehat{\theta})\right) = \mathbb{E}_{p_{\theta_0}(\cdot|\widehat{\theta})}\left[\left(1 - \frac{\frac{f(X;\widehat{\theta})}{f(X;\theta_0)}}{\mathbb{E}_{p_{\theta_0}(\cdot|\widehat{\theta})}\left[\frac{f(X;\widehat{\theta})}{f(X;\theta_0)}\right]}\right)_+\right]$$

So, we need to show that $\frac{f(X;\widehat{\theta})}{f(X;\theta_0)}$ is $\approx$ constant over distrib. $X|\widehat{\theta}$.

Proof sketch for Theorem

$$\log\left(\frac{f(X;\widehat{\theta})}{f(X;\theta_0)}\right) = -(\theta_0-\widehat{\theta})^\top \nabla_\theta \log f(X;\widehat{\theta}) - \frac{1}{2}(\theta_0-\widehat{\theta})^\top \nabla_\theta^2 \log f(X;\tilde{\theta})(\theta_0-\widehat{\theta})$$

Proof sketch for Theorem

$$\log\left(\frac{f(X;\widehat{\theta})}{f(X;\theta_0)}\right) = -(\theta_0-\widehat{\theta})^\top \nabla_\theta \log f(X;\widehat{\theta}) - \frac{1}{2}(\theta_0-\widehat{\theta})^\top \nabla_\theta^2 \log f(X;\tilde{\theta})(\theta_0-\widehat{\theta})$$

$$\implies \quad \left|\log\left(\frac{f(X;\widehat{\theta})}{f(X;\theta_0)}\right) + \frac{1}{2}(\theta_0-\widehat{\theta})^\top \mathbb{E}_{\theta_0}\left[\nabla_\theta^2 \log f(X;\tilde{\theta})\right](\theta_0-\widehat{\theta})\right|$$

$$\leq r \cdot \underbrace{\|\nabla_\theta \log f(X;\widehat{\theta})\|}_{=\sigma\|W\|\asymp\sigma} + \frac{1}{2}\cdot \underbrace{r^2 \left\|\nabla_\theta^2 \log f(X;\tilde{\theta}) - \mathbb{E}_{\theta_0}\left[\nabla_\theta^2 \log f(X;\tilde{\theta})\right]\right\|}_{\asymp\varepsilon \text{ by Asm. } 3}$$

$\|\theta_0 - \widehat{\theta}\| \leq r$
with prob. $\geq 1-\delta$ by Asm. 2

Proof sketch for Theorem

$$\log\left(\frac{f(X;\widehat{\theta})}{f(X;\theta_0)}\right) = -(\theta_0-\widehat{\theta})^\top \nabla_\theta \log f(X;\widehat{\theta}) - \frac{1}{2}(\theta_0-\widehat{\theta})^\top \nabla^2_\theta \log f(X;\tilde{\theta})(\theta_0-\widehat{\theta})$$

$$\Longrightarrow \quad \left|\log\left(\frac{f(X;\widehat{\theta})}{f(X;\theta_0)}\right) + \frac{1}{2}(\theta_0-\widehat{\theta})^\top \mathbb{E}_{\theta_0}\left[\nabla^2_\theta \log f(X;\tilde{\theta})\right](\theta_0-\widehat{\theta})\right|$$

$$\le r \cdot \underbrace{\|\nabla_\theta \log f(X;\widehat{\theta})\|}_{=\sigma\|W\|\asymp\sigma} + \frac{1}{2}\cdot \underbrace{r^2 \left\|\nabla^2_\theta \log f(X;\tilde{\theta}) - \mathbb{E}_{\theta_0}\left[\nabla^2_\theta \log f(X;\tilde{\theta})\right]\right\|}_{\asymp\varepsilon \text{ by Asm. 3}}$$

$\|\theta_0 - \widehat{\theta}\| \le r$
with prob. $\ge 1-\delta$ by Asm. 2

Rearrange $\rightsquigarrow$

$$\mathrm{d}_{\mathrm{exch}}(X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(M)}) \le \mathbb{E}_{\widehat{\theta}}\left[\mathrm{d}_{\mathrm{TV}}\big(p_{\theta_0}(\cdot|\widehat{\theta}), p_{\widehat{\theta}}(\cdot|\widehat{\theta})\big)\right] \le 3\sigma r + \delta + \varepsilon$$

Summary & open questions

- Summary: aCSS can test goodness-of-fit by
  sampling nearly-exchangeable copies of the data,
  in a much broader range of settings than CSS

Summary & open questions

- Summary: aCSS can test goodness-of-fit by
  sampling nearly-exchangeable copies of the data,
  in a much broader range of settings than CSS

- How to choose σ to balance Type I error & power?
- Connections to Bayesian methods?
- Apply to high dimensional regression / covariance estimation?
- Apply to missing data / latent variables / models with singularities?
- Extend to model-X knockoffs?

Thank you!