Zipper: Addressing Degeneracy in Algorithm-Agnostic Inference

Geng Chen, Yinxu Jia, Guanghui Wang, Changliang Zou

Nankai University

Outline

Outline

Goodness-of-Fit Testing via Predictiveness Comparison

- Due to the popularity of black box prediction methods like random forests and deep neural networks, there has been a growing interest in the so-called "*algorithm (or model)-agnostic*" inference on the *goodness-of-fit* (GoF) in regression.
- This framework aims to assess the appropriateness of a given model for prediction compared to a potentially more complex (often higher-dimensional) model.

Goodness-of-Fit Testing via Predictiveness Comparison

- Response: $Y \in \mathbb{R}$; Covariates $X \in \mathbb{R}^p$; $(Y, X) \sim P$.
- Define $\mathbb{C}(\tilde{f}, P)$ to quantify predictive capability of $\tilde{f} \in \mathcal{F}$.
- Optimal function: $f \in \arg\max_{\tilde{f} \in \mathcal{F}} \mathbb{C}(\tilde{f}, P)$.
- Examples:
  - (Negative) squared loss: $\mathbb{C}(\tilde{f}, P) = -E[\{Y - \tilde{f}(X)\}^2]$.
  - (Negative) cross-entropy loss: $\mathbb{C}(\tilde{f}, P) = E[Y \log \tilde{f}(X) + (1 - Y) \log\{1 - \tilde{f}(X)\}]$.
- GoF testing involves two classes of functions: $\mathcal{F}$ and subset $\mathcal{F}_{\mathcal{S}}$.
- Dissimilarity measure: $\psi_{\mathcal{S}} = \mathbb{C}(f, P) - \mathbb{C}(f_{\mathcal{S}}, P)$, where $f_{\mathcal{S}} \in \arg\max_{\tilde{f} \in \mathcal{F}_{\mathcal{S}}} \mathbb{C}(\tilde{f}, P)$.

$$H_0 : \psi_{\mathcal{S}} = 0 \quad \text{versus} \quad H_1 : \psi_{\mathcal{S}} > 0.$$

Goodness-of-Fit Testing via Predictiveness Comparison

- Specification Testing: Evaluates the adequacy of a class of models (e.g., parametric models) by testing if $E(Y \mid X) = g_\theta(X)$ holds almost surely. In this context, $\mathcal{F}$ is an unrestricted class, and $\mathcal{F}_\mathcal{S}$ represents parametric models.
- Model Selection: Used to identify the superior predictive model from candidates, often comparing an unregularized model to a regularized one. Testing H_0 assesses if a regularizer improves predictions.
- Variable Importance Measure: Evaluates the significance of a covariate group U in predicting the response Y, with $X = (U^\top, V^\top)^\top$. This can be expressed in the GoF framework by defining $\mathcal{F}_\mathcal{S}$ to exclude U.

The Degeneracy Issue

The null hypothesis $H_0 : \psi_{\mathcal{S}} = 0$ poses challenges due to degeneracy (Verdinelli and Wasserman, 2024; Dai et al., 2024; Williamson et al., 2023).

- Consider testing if $\mu := E(Y) = 0$ with $\mathcal{F} = \mathbb{R}$ and $\mathcal{F}_{\mathcal{S}} = \{0\}$.
  Using squared loss,
  $\psi_{\mathcal{S}} = E(Y^2) - E\{(Y - \mu)^2\} = \mu^2$.
- The estimator based on sample-splitting is
  $\psi_{n,\mathcal{S}} = 2\bar{Y}_n^{\mathrm{te}}\bar{Y}_n^{\mathrm{tr}} - (\bar{Y}_n^{\mathrm{tr}})^2$.
- When $\mu \neq 0$, $\sqrt{n}(\psi_{n,\mathcal{S}} - \mu)$ is asymptotically normal.
  However, under H_0, $\sqrt{n}\psi_{n,\mathcal{S}} = O_P(n^{-1/2})$, indicating degeneracy.
- While inference at a n-rate is feasible in this simple case, degeneracy poses challenges for more complex models and black box algorithms.

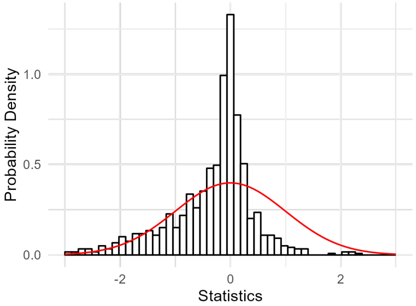


Figure: Empirical distribution of $\sqrt{n}\psi_{n,\mathcal{S}}$ scaled by its standard deviation (black histograms) compared to normal distribution (red lines).

Existing Solutions

- Sample Splitting: Williamson et al. (2023) additionally split the testing data to evaluate the nondegenerate influence functions of $\mathbb{C}(f, P)$ and $\mathbb{C}(f_{\mathcal{S}}, P)$ separately under H_0. However, this reduce sample size and significantly lower power.
- Data Perturbation: Rinaldo et al. (2019) and Dai et al. (2024) proposed adding independent zero-mean noise to empirical influence functions. However, determining the right amount of perturbation remains a heuristic process.
- Standard Error Expansion: Verdinelli and Wasserman (2024) suggested expanding the standard error of the estimator to mitigate the effects of degeneracy.

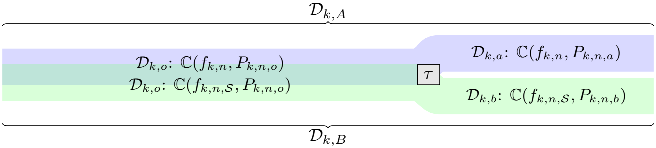
Our Contributions

- We introduce the Zipper device for algorithm-agnostic inference under the null hypothesis H_0 of equal goodness.
- Our approach utilizes overlapping testing splits with a *slider* parameter $\tau \in [0, 1)$, enhancing data efficiency and significantly improving power while ensuring valid size control.

Outline

The Zipper Device

- Randomly partition data into K folds, $\mathcal{D}_1, \ldots, \mathcal{D}_K$, with estimators $f_{k,n}$ and $f_{k,n,\mathcal{S}}$ for f and $f_{\mathcal{S}}$ constructed from data excluding fold $\mathcal{D}_k$.
- Split $\mathcal{D}_k$ into two overlapping sets $\mathcal{D}_{k,A}$ and $\mathcal{D}_{k,B}$, adjusting the overlap proportion through $\tau = |\mathcal{D}_{k,o}|/|\mathcal{D}_{k,A}|$.



- Construct estimators $\mathbb{C}_{k,n}$ and $\mathbb{C}_{k,n,\mathcal{S}}$ for $\mathbb{C}(f, P)$ and $\mathbb{C}(f_{\mathcal{S}}, P)$ using $(f_{k,n}, \mathcal{D}_{k,A})$ and $(f_{k,n,\mathcal{S}}, \mathcal{D}_{k,B})$.
- The estimator of $\psi_{\mathcal{S}}$ is $\psi_{n,\mathcal{S}} = K^{-1} \sum_{k=1}^{K} (\mathbb{C}_{k,n} - \mathbb{C}_{k,n,\mathcal{S}})$.

The Zipper Device

- $\tau = 0$: aligns the vanilla sample splitting method (Williamson et al., 2023; Dai et al., 2024).
- $\tau = 1$: $\mathcal{D}_{k,o} = \mathcal{D}_{k,A} = \mathcal{D}_{k,B} = \mathcal{D}_k$, leading to the degeneracy under H_0 (Williamson et al., 2023).
- Restrict the slider parameter $\tau \in [0, 1)$.

Asymptotic Linearity

Theorem (Asymptotic linearity)

If Conditions (C1)–(C4) hold for both tuples $(P, \mathcal{F}, f, f_{k,n})$ *and* $(P, \mathcal{F}_{\mathcal{S}}, f_{\mathcal{S}}, f_{k,n,\mathcal{S}})$, *then*

$$\psi_{n,\mathcal{S}} - \psi_{\mathcal{S}} = \frac{1}{n/(2-\tau)} \sum_{k=1}^{K} \Bigg[\sum_{i:Z_i \in \mathcal{D}_{k,a}} \phi(Z_i) - \sum_{i:Z_i \in \mathcal{D}_{k,b}} \phi_{\mathcal{S}}(Z_i) + \sum_{i:Z_i \in \mathcal{D}_{k,o}} \{\phi(Z_i) - \phi_{\mathcal{S}}(Z_i)\} \Bigg] + o_P(n^{-1/2}),$$

where $\phi(Z) = \dot{\mathbb{C}}(f, P; \delta_Z - P)$ *and* $\phi_{\mathcal{S}}(Z) = \dot{\mathbb{C}}(f_{\mathcal{S}}, P; \delta_Z - P)$. *Here,* $\dot{\mathbb{C}}(\tilde{f}, P; h)$ *represents the Gâteaux derivative of* $\tilde{P} \mapsto \mathbb{C}(\tilde{f}, \tilde{P})$ *at* P *in the direction* h, *and* δ_z *denotes the Dirac measure at* z. *Consequently, for any* $\tau \in [0, 1)$,

$$\{n/(2-\tau)\}^{1/2}(\psi_{n,\mathcal{S}} - \psi_{\mathcal{S}}) \xrightarrow{d} N(0, \nu^2_{\mathcal{S},\tau})$$

as $n \to \infty$, *where* $\nu^2_{\mathcal{S},\tau} = (1-\tau)(\sigma^2 + \sigma^2_{\mathcal{S}}) + \tau\eta^2_{\mathcal{S}}$, $\sigma^2 = E\{\phi^2(Z)\}$, $\sigma^2_{\mathcal{S}} = E\{\phi^2_{\mathcal{S}}(Z)\}$, *and* $\eta^2_{\mathcal{S}} = E[\{\phi(Z) - \phi_{\mathcal{S}}(Z)\}^2]$.

Null Behaviors

- Use the plug-in principle to obtain the variance estimator $\nu^2_{n,\mathcal{S},\tau}$ for $\nu^2_{\mathcal{S},\tau}$ under H_0.
  - $\nu^2_{n,\mathcal{S},\tau}$ converges to $\nu^2_{\mathcal{S},\tau}$ as $n \to \infty$ under H_0 if Conditions (C4)–(C5) are satisfied.
- The normalized test statistic is given by

$$T_\tau = \frac{\sqrt{n/(2-\tau)}\psi_{n,\mathcal{S}}}{\nu_{n,\mathcal{S},\tau}}.$$

  Reject H_0 if $T_\tau > z_{1-\alpha}$ for a prespecified significance level α.
- Under H_0, since $T_\tau \xrightarrow{d} N(0,1)$, Zipper ensures valid size control.

Power Analysis

Theorem (Power approximation)

Suppose the Conditions (C1)–(C5) hold for both tuples $(P, \mathcal{F}, f, f_{k,n})$ *and* $(P, \mathcal{F}_{\mathcal{S}}, f_{\mathcal{S}}, f_{k,n,\mathcal{S}})$. *Then for any* $\tau \in [0, 1)$, *the power function* $\Pr(T_\tau > z_{1-\alpha} \mid H_1) = G_{\mathcal{S},n,\alpha}(\tau) + o(1)$, *where*

$$G_{\mathcal{S},n,\alpha}(\tau) = \Phi\left(-\frac{\nu^{(0)}_{\mathcal{S},\tau}}{\nu_{\mathcal{S},\tau}} z_{1-\alpha} + \frac{\{n/(2-\tau)\}^{1/2}\psi_{\mathcal{S}}}{\nu_{\mathcal{S},\tau}}\right),$$

$\nu^{(0)}_{\mathcal{S},\tau} = \{(1-\tau)(\sigma^2 + \sigma^2_{\mathcal{S}})\}^{1/2}$ *and* Φ *denotes the distribution function of* $N(0, 1)$. *Furthermore, if* $Cov\{\phi(Z), \phi_{\mathcal{S}}(Z)\} \geq 0$, *then* $G_{\mathcal{S},n,\alpha}(\tau)$ *increases with* τ.

Power Analysis

- Sample Splitting: At $\tau = 0$, the approximate power function is:

$$G_{\mathcal{S},n,\alpha}(0) = \Phi\left(-z_{1-\alpha} + \frac{(n/2)^{1/2}\psi_{\mathcal{S}}}{(\sigma^2 + \sigma_{\mathcal{S}}^2)^{1/2}}\right).$$

- Zipper: For $\tau \in [0, 1)$, power function satisfies:

$$G_{\mathcal{S},n,\alpha}(\tau) \overset{(i)}{\geq} \Phi\left(-z_{1-\alpha} + \frac{\{n/(2-\tau)\}^{1/2}\psi_{\mathcal{S}}}{(\sigma^2 + \sigma_{\mathcal{S}}^2)^{1/2}}\right) \overset{(ii)}{\geq} G_{\mathcal{S},n,\alpha}(0).$$

- The power improvement of Zipper compared to sample splitting comes from
  - the introduction of overlap mechanism τ (Inequality (ii)).
  - the utilization of variance estimator $\nu_{n,\mathcal{S},\tau}^2$ (Inequality (ii)).

Efficiency-and-Degeneracy Tradeoff

- To achieve better power while maintaining a reliable size, we propose a simple approach for selecting τ.
- To ensure a favorable normal approximation, we can choose the sample size $(1-\tau)n/(2-\tau)$ such that it meets a predetermined "large" sample size, such as $n_0 = 30$ or 50. Say, we can specify $\tau = \tau_0 := (n - 2n_0)/(n - n_0)$.
- In the case of very large n, a truncation may be needed to safeguard against degeneracy. For example, we can set $\tau = \min\{\tau_0, 0.9\}$.

Outline

Variable Importance Assessment

- Models Considered:
  - Normal Response: $Y \sim N(X^\top \beta, \sigma_Y^2)$.
  - Binomial Response: $Y \sim \mathrm{binom}(1, \mathrm{logit}(X^\top \beta))$.
- Design Scenarios: $n = 500$.
  - Low-Dimensional: $p = 5$ with $\beta = (\delta, \delta, 5, 0, 5, 0_{p-5})^\top$.
  - High-Dimensional: $p = 1000$ with $\beta = (\delta, \delta, 5_{0.01p}, 0^\top_{0.99p-2})^\top$.
- Test the significance of the first two variables given the significance level $\alpha = 5\%$.
- $\tau = \min\{\tau_0, 0.9\}$ with $n_0 = 50$.

Table: Empirical sizes (in percentage) of various testing procedures, with standard deviations in brackets.

Model	p	Zipper	WGSC-3	DSP-Split	WGSC-2	DSP-Pert
Normal	5	3.9(0.19)	5.1(0.22)	4.6(0.21)	0.1(0.03)	10.2(0.30)
	1000	4.3(0.20)	6.2(0.24)	5.9(0.24)	16.7(0.37)	35.0(0.48)
Binomial	5	3.7(0.19)	3.9(0.19)	4.2(0.20)	0.6(0.08)	4.0(0.20)
	1000	5.6(0.23)	4.8(0.21)	5.1(0.22)	19.9(0.40)	38.6(0.49)

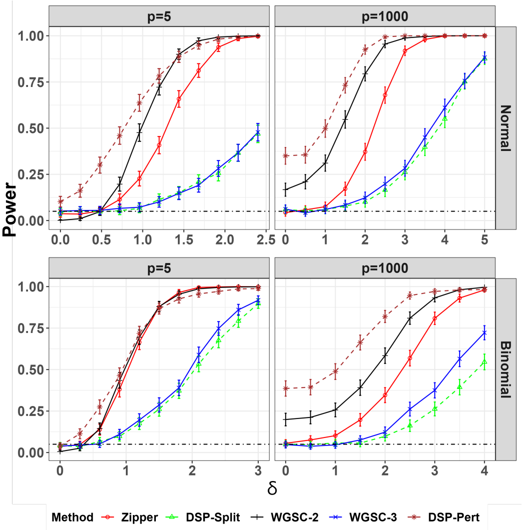


Figure: Empirical power of various testing methods as a function of the magnitude δ. The dot-dashed horizontal line represents the intercept at $\alpha = 5\%$.

Model Specification Testing

- $Y = X\beta + \varepsilon$, where $\|\beta\|_0 = 2$.
- $H_0 : \beta = (*, *, 0_{p-2})^\top$ vs $H_1 : \|\beta\|_0 = 2$ but not H_0 (with $*$ as any nonzero value).
- Scenarios:
  - (i) $\beta = (0.4, 0.4, 0_{p-2})^\top$ (under H_0).
  - (ii) $\beta = (0.4, 0, 0.4, 0_{p-3})^\top$ (under H_1).
  - (iii) $\beta = (0, 0, 0.4, 0.4, 0_{p-4})^\top$ (under H_1).
- Estimation Methods:
  - Best subset selection for $p = 5$.
  - Abess (Zhu et al. (2022)) for $p = 1000$.

Table: Empirical sizes and powers (in percentage) for the model specification test, with standard deviations in brackets.

p	5				1000			
Scenerio	Zipper	WGSC-3	DSP-Split	WGSC-2	Zipper	WGSC-3	DSP-Split	WGSC-2
(i)	4.3(0.20)	6.2(0.22)	5.6(0.20)	0.0(0.00)	4.2(0.19)	5.5(0.20)	6.5(0.21)	16.6(0.36)
(ii)	96.9(0.17)	31.2(0.46)	34.9(0.46)	100.0(0.00)	94.2(0.22)	29.8(0.46)	31.4(0.46)	97.3(0.16)
(iii)	100.0(0.00)	81.4(0.39)	79.3(0.38)	100.0(0.00)	100.0(0.00)	81.3(0.40)	78.1(0.41)	100.0(0.00)

MNIST Handwritten Dataset

- MNIST dataset consists of size-normalized and center-aligned handwritten digit images. Each image is represented as a 28×28 pixel grid ($p = 28^2 = 784$).
- Focused on digits 7 and 9, resulting in $n = 14251$ images.
- Images divided into nine distinct regions. Conduct variable importance testing for each region while considering others.
- Employed a Convolutional Neural Network (CNN) for image analysis.
- Set significance level for tests at $\alpha = 0.05/9$ using Bonferroni correction.

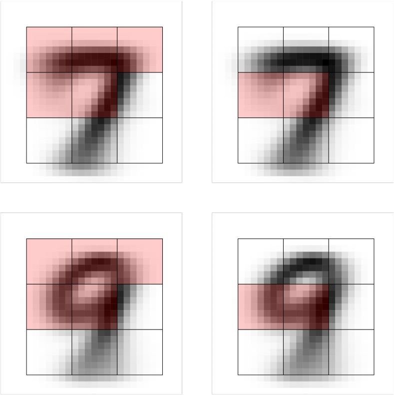


Figure: Hypothesis regions (blank squares) and important discoveries (squares filled in red) comparing the Zipper method (left column) with WGSC-3 (right column).

Bodyfat Dataset

- The bodyfat dataset (Penrose et al., 1985) provides an estimate of body fat percentages obtained through underwater weighing, along with various body circumference measurements from a sample of $n = 252$ men.
- Conduct variable importance tests for each body circumference while considering potential influences from essential attributes such as age, weight, and height.
- Employ the random forest for accurate regression function estimation.
- Set significance level for tests at $\alpha = 0.05/10$ using Bonferroni correction.

Table: P-values obtained from the Zipper and WGSC-3 methods for each marginal test regarding the relevance of the body circumference.

Body Part	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
Zipper	0.98	0.10	5.48×10^{-10}	4.01×10^{-4}	0.10	0.03	0.20	0.26	0.35	0.02
WGSC-3	0.12	0.01	9.30×10^{-4}	0.29	0.01	0.06	0.36	0.18	0.69	0.05

- The Zipper method identifies both Abdomen and Hip as significant factors. In contrast, WGSC-3 suggests only Abdomen as important.
- A recent study by Zhu et al. (2023) proposed the formula $(\text{Waist} + \text{Hip})/\text{Height}$ as a straightforward body fat evaluation index, which aligns with our findings.

Outline

Concluding Remarks

- We introduce Zipper, an effective tool for addressing degeneracy in algorithm/model-agnostic inference.
- The mechanism of Zipper involves the recycling of data usage by constructing two overlapping data splits within the testing samples, which holds potential for independent exploration.
- Furthermore, incorporating the Zipper device into large-scale comparisons to achieve error rate control warrants additional research.

Reference

Dai, B., Shen, X., and Pan, W. (2024). Significance tests of feature relevance for a black-box learner. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):1898–1911.

Penrose, K. W., Nelson, A., and Fisher, A. (1985). Generalized body composition prediction equation for men using simple measurement techniques. *Medicine & Science in Sports & Exercise*, 17(2):189.

Rinaldo, A., Wasserman, L., and G'Sell, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, 47(6):3438–3469.

Verdinelli, I. and Wasserman, L. (2024). Decorrelated variable importance. *Journal of Machine Learning Research*, 25(7):1–27.

Williamson, B. D., Gilbert, P. B., Simon, N., and Carone, M. (2023). A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 118(543):1645–1658.

Zhu, J., Wang, X., Hu, L., Huang, J., Jiang, K., Zhang, Y., Lin, S., and Zhu, J. (2022). abess: A fast best-subset selection library in python and r. *Journal of Machine Learning Research*, 23(202):1–7.

Zhu, Y., Maruyama, H., Onoda, K., Zhou, Y., Huang, Q., Hu, C., Ye, Z., Li, B., and Wang, Z. (2023). Body mass index combined with (waist + hip)/height accurately screened for normal-weight obesity in chinese young adults. *Nutrition*, 108:111939.