UNIVERSITY of WASHINGTON

The Rao, Wald, and Likelihood-Ratio Tests under Generalized Self-Concordance

Lang Liu, Zaid Harchaoui

Department of Statistics, University of Washington

Overview

- We recover the **asymptotic equivalence** of the Rao, Wald, and likelihood-ratio tests from a **non-asymptotic viewpoint**.
- We characterize the **critical sample size** beyond which the equivalence holds asymptotically under the **null** hypotheses.
- We also analyze the **statistical power** under both the **fixed and alternative** hypotheses.
- We establish an estimation bound that matches the **misspecified Cramér-Rao** lower bound.

Goodness-of-Fit Testing

Problem. Let $Z \sim \mathbb{P}$ and $\mathcal{P}_\Theta := \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$. Assume there exists $\theta_\star \in \Theta$ such that $\mathbb{P} = P_{\theta_\star}$. Given an i.i.d. sample $\{Z_i\}_{i=1}^n$, we want to infer properties of $\theta_\star$ via

$$\mathcal{H}_0 : \theta_\star = \theta_0 \leftrightarrow \mathcal{H}_1 : \theta_\star \neq \theta_0.$$

- A **test statistic** $T := T(Z_1, \dots, Z_n)$ and a **prescribed critical value** t_n.
- **Reject** the null $\mathcal{H}_0$ if $T > t_n$.
- **Type I error rate** $\Pr(T > t_n \mid \mathcal{H}_0)$ and **statistical power** $\Pr(T > t_n \mid \mathcal{H}_1)$.

Notation. Loss function $\ell(\theta; z) := -\log P_\theta(z)$.

- **Population risk** $L(\theta) := \mathbb{E}[-\log P_\theta(Z)]$ and **empirical risk** $L_n(\theta) := \frac{1}{n}\sum_{i=1}^n \ell(\theta; Z_i)$.
- **Empirical risk minimizer** $\theta_n := \arg\min_{\theta\in\Theta} L_n(\theta)$.
- **Gradient** $S(\theta; z) := \nabla_\theta \ell(\theta; z)$, $S(\theta) := \mathbb{E}[S(\theta; Z)] = \nabla_\theta L(\theta)$, and $S_n(\theta) := \frac{1}{n}\sum_{i=1}^n S(\theta; Z_i)$.
- **Hessian** $H(\theta; z) := \nabla^2_\theta \ell(\theta; z)$, $H(\theta) := \mathbb{E}[H(\theta; Z)]$, and $H_n(\theta) := \frac{1}{n}\sum_{i=1}^n H(\theta; Z_i)$.

Three classical goodness-of-fit tests.

- **Rao test**—$T_{Rao} := S_n(\theta_0)^\top H_n(\theta_0)^{-1} S_n(\theta_0)$.
- **Wald test**—$T_{Wald} := (\theta_n - \theta_0)^\top H_n(\theta_n)(\theta_n - \theta_0)$.
- **The likelihood-ratio test**—$T_{LR} := 2[L_n(\theta_0) - L_n(\theta_n)]$.

Asymptotic equivalence of the three tests.

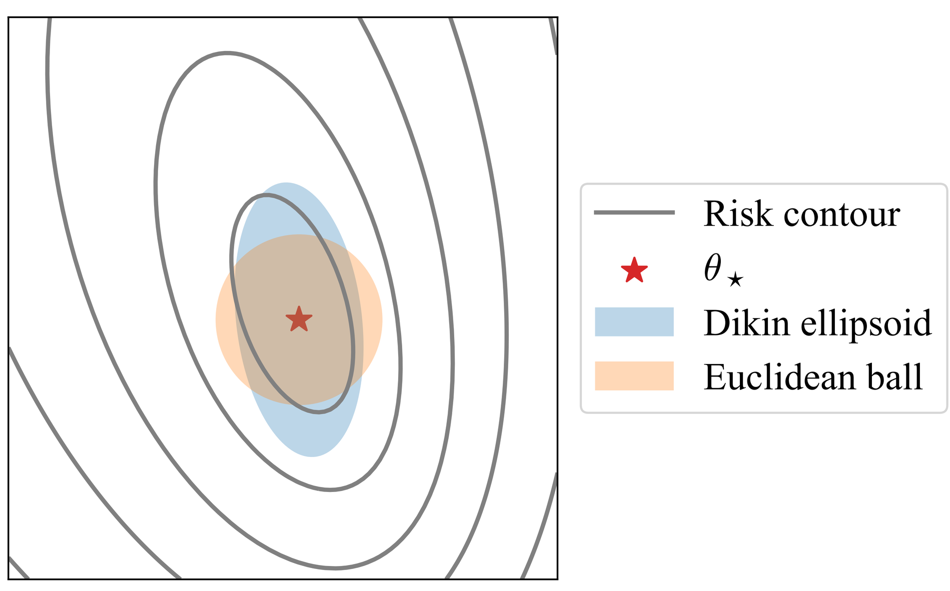
- Asymptotic distribution.
  - $\sqrt{n} S_n(\theta_0) \to_d \mathcal{N}_d(0, G(\theta_0))$ under $\mathcal{H}_0$.
  - Well-specified model, i.e., $\mathbb{P} \in \mathcal{P}_\Theta$, implies that $G(\theta_0) = H(\theta_0)$.
  - $nT_{Rao} \to_d \chi^2_d$ under $\mathcal{H}_0$.
- Asymptotic equivalence.
  - $S_n(\theta_0) = S_n(\theta_0) - S_n(\theta_n) = H_n(\bar\theta_n)(\theta_0 - \theta_n)$.
  - $T_{Rao} = (\theta_n - \theta_0)^\top H_n(\theta_n)(\theta_n - \theta_0) + o_p(1) = T_{Wald} + o_p(1)$.
  - $T_{LR} = 2S_n(\theta_n)(\theta_0 - \theta_n) + (\theta_0 - \theta_n)^\top H_n(\bar\theta_n)(\theta_0 - \theta_n) = T_{Wald} + o_p(1)$.

Prelimilaries

Dikin ellipsoid. A Dikin ellipsoid at $\theta_\star$ of radius r

$$\Theta_r(\theta_\star) := \{\theta \in \Theta : \|H_\star^{1/2}(\theta - \theta_\star)\| < r\}.$$

- The shape of a **Euclidean ball** is **always the same**.
- The shape of a **Dikin ellipsoid** is **adapted to the geometry** near the optimum.



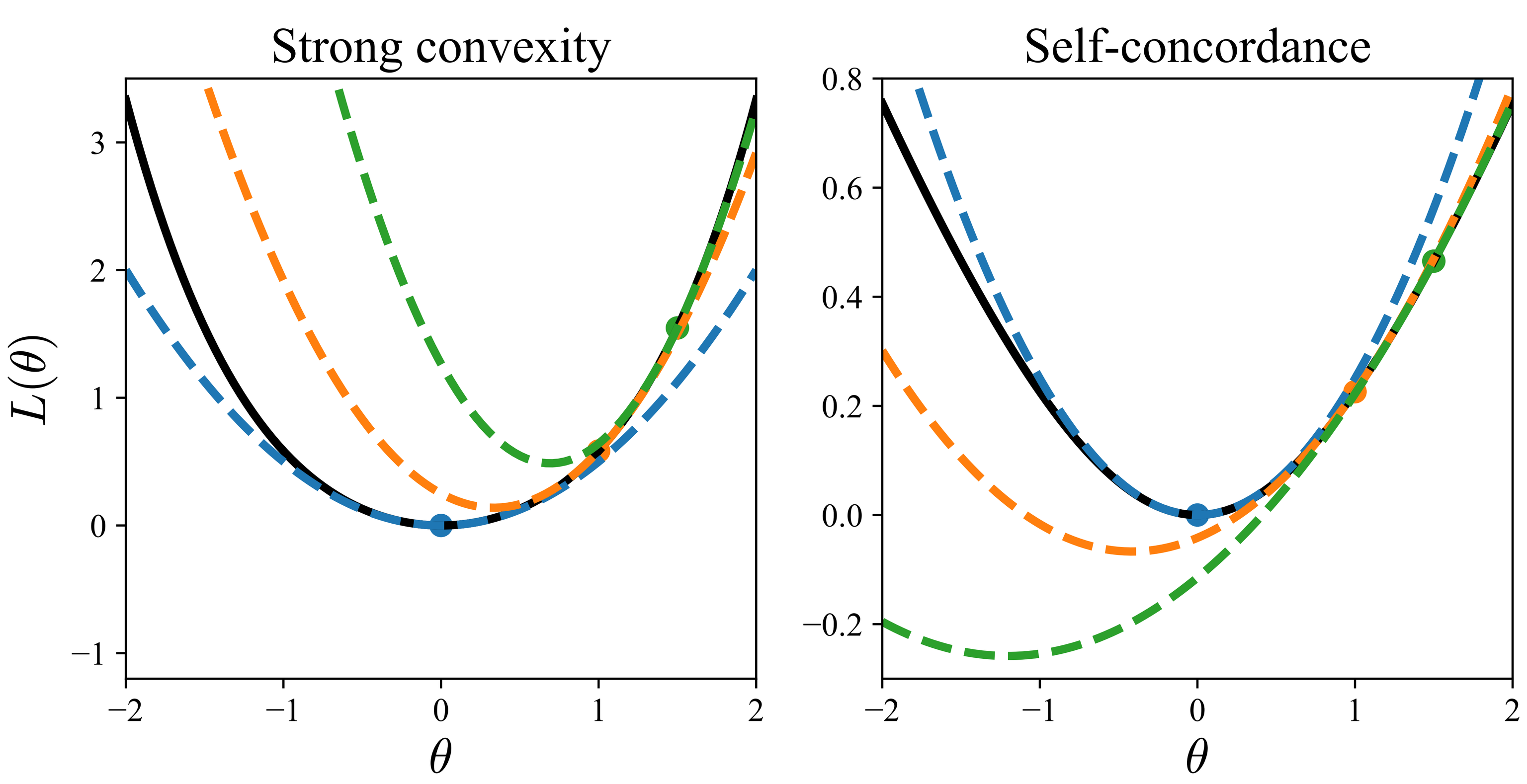
Generalized self-concordance. Let f be convex, $R > 0$, and $\nu > 0$. We say f is (R, ν)-generalized self-concordant if (on a high level)

$$\left\|\nabla^3 f(x)\right\| \lesssim R \left\|\nabla^2 f(x)\right\|^\nu.$$

We give two examples of losses as functions of parameters.

- For **linear regression**, its loss is (R, ν)-generalized self-concordant for any $R > 0$ and $\nu > 0$.
- For **logistic regression** with $\|X\|_2 \le_{a.s.} M$, its loss is $(2M, 2)$-generalized self-concordant.

	Strong convexity	Self-concordance
Hessian lower bound	Global	Local
Hessian varying rate	No control	Slow



Concentration of Hessian. A key result towards deriving our bounds is

$$[1 - c_n(\delta)]H(\theta) \preceq H_n(\theta) \preceq [1 + c_n(\delta)]H(\theta)$$

with probability at least $1 - \delta$, where $c_n(\delta) = O(\sqrt{\log(d/\delta)/n})$.

Main Results

Type I error rate. Let $\theta_\star = \theta_0$. We have, with probability at least $1 - \delta$,

$$nT_{Rao} \lesssim d + \log\frac{e}{\delta}, \quad \text{whenever } n \gtrsim \log\frac{2d}{\delta}.$$

Additionally, with $\lambda_\star := \lambda_{\min}(H_\star)$,

$$nT_{Wald}, nT_{LR} \lesssim d + \log\frac{e}{\delta}, \quad \text{whenever } n \gtrsim \log\frac{2d}{\delta} + d\frac{R^2}{\lambda_\star^{3-\nu}}.$$

- Demonstrate that the three tests have a tail behavior that is governed by a χ^2_d distribution.
- Characterize the **critical sample size** enough to enter the asymptotic regime.

Statistical power. Let $\theta_\star \to_{n\to\infty} \theta_0$. Let $t_n(\alpha)$ be the $(1-\alpha)$-quantile of χ^2_d. Let $\Omega(\theta) := G(\theta)^{\frac{1}{2}} H(\theta)^{-1} G(\theta)^{\frac{1}{2}}$ and $h(\tau) := \min\{\tau^2, \tau\}$. The following statements hold for sufficiently large n.

- Let $\tau_n \asymp \left\|H(\theta_0)^{1/2}(\theta_\star - \theta_0)\right\|^2$. We have

$$\Pr(T_{Rao} > t_n(\alpha)) \begin{cases} \le 2de^{-Cn} + e^{-Ch(\|\Omega(\theta_0)\|_2^{-1})} & \text{if } \left\|H(\theta_0)^{1/2}(\theta_\star - \theta_0)\right\| = O(n^{-1/2}) \\ \ge 1 - 2de^{-Cn} - e^{-Ch(n\tau_n\|\Omega(\theta_0)\|_2^{-1})} & \text{if } \left\|H(\theta_0)^{1/2}(\theta_\star - \theta_0)\right\| = \omega(n^{-1/2}). \end{cases}$$

- Let $\tau_n' \asymp \left\|H(\theta_0)^{1/2}(\theta_\star - \theta_0)\right\|^2$. We have

$$\Pr(T_{Wald}, T_{LR} > t_n(\alpha)) \begin{cases} \le 2nde^{-C(\frac{\lambda_\star^{3-\nu} n}{R^2 d})^{\frac{1}{\nu-1}}} + e^{-Ch(\|\Omega(\theta_\star)\|_2^{-1})} & \text{if } \left\|H(\theta_0)^{1/2}(\theta_\star - \theta_0)\right\| = O(n^{-1/2}) \\ \ge 1 - 2nde^{-C(\frac{\lambda_\star^{3-\nu} n}{R^2 d})^{\frac{1}{\nu-1}}} - e^{-Ch(n\tau_n'\|\Omega(\theta_\star)\|_2^{-1})} & \text{if } \left\|H(\theta_0)^{1/2}(\theta_\star - \theta_0)\right\| = \omega(n^{-1/2}). \end{cases}$$

To summarize,

- If $\left\|H(\theta_0)^{1/2}(\theta_\star - \theta_0)\right\| = O(n^{-1/2})$, the power is asymptotically **upper bounded by a constant**.
- If $\left\|H(\theta_0)^{1/2}(\theta_\star - \theta_0)\right\| = \omega(n^{-1/2})$, the power **tends to one** at rate $O\big(\exp(-n\|H(\theta_0)^{1/2}(\theta_\star - \theta_0)\|^2)\big)$.

Estimation Bound under Model Misspecification

Estimation bound. Assume $\mathbb{P} \notin \mathcal{P}_\Theta$ and let $\theta_\star := \arg\min_{\theta\in\Theta} L(\theta)$. It holds that

$$\left\|H_n(\theta_n)^{1/2}(\theta_n - \theta_\star)\right\|^2 \lesssim \frac{d_\star}{n} + \frac{\|\Omega(\theta_\star)\|_2}{n}\log\frac{e}{\delta},$$

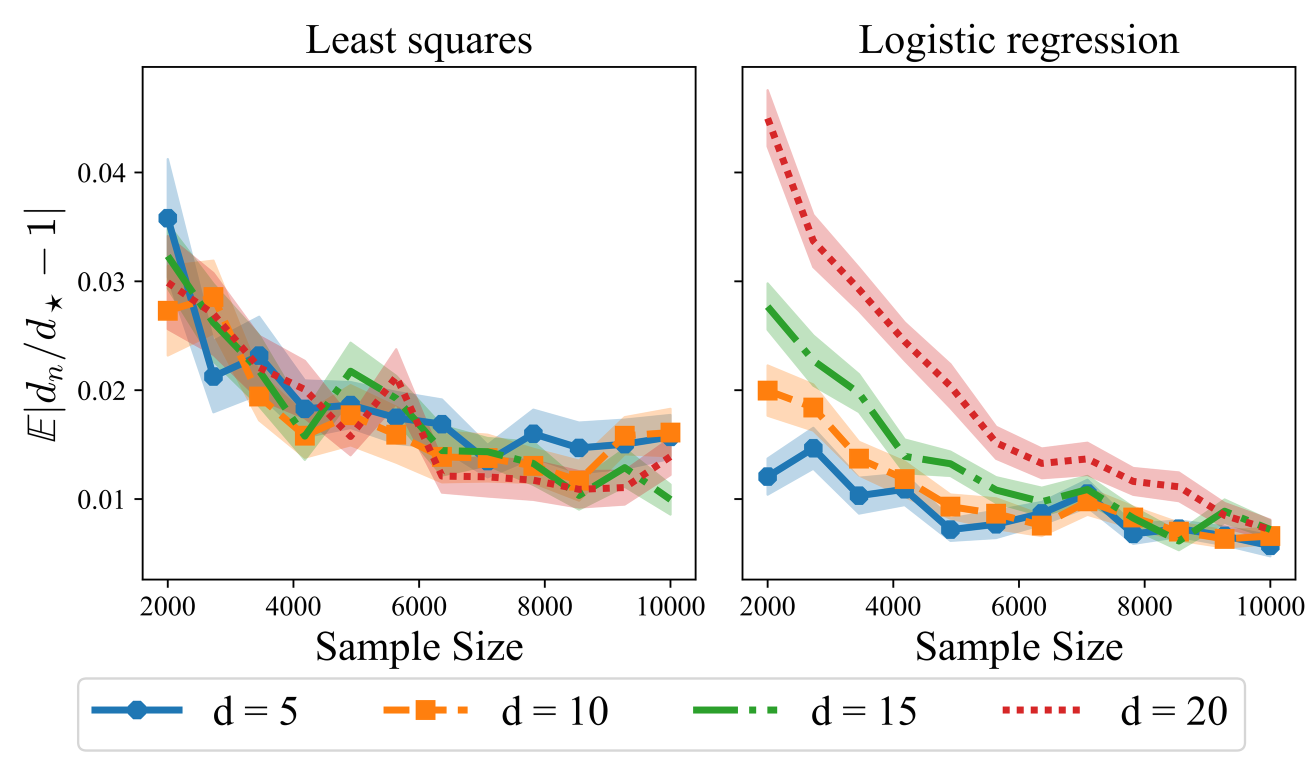
where $d_\star := \mathbf{Tr}(\Omega(\theta_\star))$ is the **effective dimension**.

- When the model is **well-specified**, $d_\star = d$ and $\|\Omega(\theta_\star)\|_2 = 1$.
- When the model is **misspecified**,

	Eigendecay		Dimension Dependency		Ratio
	$G_\star$	$H_\star$	$d_\star$	d	$d_\star/d$
Poly-Poly	$i^{-\alpha}$	$i^{-\beta}$	$d^{(\beta-\alpha+1)\vee 0}$	d	$d^{(\beta-\alpha)\vee(-1)}$
Poly-Exp	$i^{-\alpha}$	$e^{-\nu i}$	$d^{1-\alpha}e^{\nu d}$	d	$d^{-\alpha}e^{\nu d}$
Exp-Poly	$e^{-\mu i}$	$i^{-\beta}$	1	d	d^{-1}
Exp-Exp	$e^{-\mu i}$	$e^{-\nu i}$	d if $\mu = \nu$; 1 if $\mu > \nu$; $e^{(\nu-\mu)d}$ if $\mu < \nu$	d	1 if $\mu = \nu$; d^{-1} if $\mu > \nu$; $d^{-1}e^{(\nu-\mu)d}$ if $\mu < \nu$

The bound matches the **Cramér-Rao** lower bound.

- $d_\star$ can be approximated by $d_n := \mathbf{Tr}(G_n(\theta_n)^{1/2} H_n(\theta_n)^{-1} G_n(\theta_n)^{1/2})$.
- **How well does d_n approximate $d_\star$?**



Examples

Generalized linear models. Consider the statistical model

$$p_\theta(y \mid x) \sim \frac{\exp\left[\theta^\top t(x,y) + h(x,y)\right]}{\int \exp\left[\theta^\top t(x,\bar y) + h(x,\bar y)\right]\mathrm{d}\mu(\bar y)}\mathrm{d}\mu(y)$$

with $\|t(X,Y)\|_2 \le_{a.s.} M$. It induces the loss function

$$\ell(\theta; z) := -\theta^\top t(x,y) - h(x,y) + \log\int \exp\left[\theta^\top t(x,\bar y) + h(x,\bar y)\right]\mathrm{d}\mu(\bar y),$$

which is $(2M, 2)$-generalized self-concordant.

Score matching with exponential families. Consider an exponential family with density $\log p_\theta(z) = \theta^\top t(z) + h(z) - \Lambda(\theta)$. The score matching loss is

$$\ell(\theta; z) = \frac{1}{2}\theta^\top A(z)\theta - b(z)^\top\theta + c(z) + \mathit{const},$$

where $A(z) := \sum_{k=1}^p \frac{\partial t(z)}{\partial z_k}\left(\frac{\partial t(z)}{\partial z_k}\right)^\top$ is positive semi-definite,

$$b(z) := \sum_{k=1}^p \left[\frac{\partial^2 t(z)}{\partial z_k^2} + \frac{\partial h(z)}{\partial z_k}\frac{\partial t(z)}{\partial z_k}\right], \text{ and } c(z) := \sum_{k=1}^p \left[\frac{\partial^2 h(z)}{\partial z_k^2} + \left(\frac{\partial h(z)}{\partial z_k}\right)^2\right].$$

It is generalized self-concordant for all $\nu \ge 2$ and $R \ge 0$.