A Kernelized Stein Discrepancy for Goodness-of-fit Tests

Qiang Liu, Dartmouth College

Jason D. Lee and Michael I. Jordan University of California, Berkeley

September 3, 2016

Introduction

- Goodness-of-fit (GOF) tests: Given a distribution p and observation $\{x_i\}$ (drawn from unknown q), test

$$H_0 : \{x_i\} \text{ is drawn from } p \quad (\text{or} \ \ p = q)$$

- Motivation: checking model assumptions, model evaluation, etc.
- We are interested in complex, high dimensional distributions $p(x)$, often with intractable normalization constants.

$$p(x) = \frac{1}{Z}\bar{p}(x), \qquad \textbf{Normalization constant: } Z = \int \bar{p}(x)dx.$$

e.g., graphical models, (restricted) Boltzmann machines, etc

Z: often critically difficult to calculate

Introduction

- Goodness-of-fit (GOF) tests: Given a distribution p and observation $\{x_i\}$ (drawn from unknown q), test

$$H_0 : \{x_i\} \text{ is drawn from } p \quad (\text{or } \; p = q)$$

- Motivation: checking model assumptions, model evaluation, etc.
- We are interested in complex, high dimensional distributions $p(x)$, often with intractable normalization constants.

$$p(x) = \frac{1}{Z}\bar{p}(x), \qquad \textbf{Normalization constant: } Z = \int \bar{p}(x)dx.$$

e.g., graphical models, (restricted) Boltzmann machines, etc

Z: often critically difficult to calculate

Introduction

- Goodness-of-fit (GOF) tests: Given a distribution p and observation $\{x_i\}$ (drawn from unknown q), test

$$H_0: \; p = q \quad vs. \quad p \neq q$$

- Challenges:
  1. Classical GOF tests, such as chi-square, Kolmogorov-Smirnov, only works for simple, low dimensional distributions.
  2. We can simulate $\{y_i\} \sim p$ and perform two-sample tests (e.g., by maximum mean discrepancy (MMD)): would not work when it is intractable to draw sample from p (MCMC may be needed).

Introduction

- Goodness-of-fit (GOF) tests: Given a distribution p and observation $\{x_i\}$ (drawn from unknown q), test

$$H_0: \; p = q \quad vs. \quad p \neq q$$

- Challenges:
  1. Classical GOF tests, such as chi-square, Kolmogorov-Smirnov, only works for simple, low dimensional distributions.
  2. We can simulate $\{y_i\} \sim p$ and perform two-sample tests (e.g., by maximum mean discrepancy (MMD)): would not work when it is intractable to draw sample from p (MCMC may be needed).

Introduction

- Goodness-of-fit (GOF) tests: Given a distribution p and observation $\{x_i\}$ (drawn from unknown q), test

$$H_0: \; p = q \quad vs. \quad p \neq q$$

- Challenges:
  1. Classical GOF tests, such as chi-square, Kolmogorov-Smirnov, only works for simple, low dimensional distributions.
  2. We can simulate $\{y_i\} \sim p$ and perform two-sample tests (e.g., by maximum mean discrepancy (MMD)): would not work when it is intractable to draw sample from p (MCMC may be needed).

Stein's Method [Stein, 1972]

- A general theoretical tool for bounding differences between distributions
  - Mostly used for theoretical proof: central limit theorem, concentration inequalities, etc.
- Key idea: Characterizing a distribution p with a Stein operator $\mathcal{A}_p$, such that

$$p = q \quad \iff \quad \mathbb{E}_{x\sim q}[\mathcal{A}_p f(x)] = 0.$$

  - For continuous distributions with smooth density $p(x)$,

$$\mathcal{A}_p f(x) \stackrel{def}{=} \nabla_x \log p(x) \cdot f(x)^\top + \nabla_x f(x).$$

Stein's Method [Stein, 1972]

- A general theoretical tool for bounding differences between distributions
  - Mostly used for theoretical proof: central limit theorem, concentration inequalities, etc.
- Key idea: Characterizing a distribution p with a Stein operator $\mathcal{A}_p$, such that

$$p = q \quad \iff \quad \mathbb{E}_{x\sim q}[\mathcal{A}_p f(x)] = 0.$$

  - For continuous distributions with smooth density $p(x)$,

$$\mathcal{A}_p f(x) \overset{def}{=} \nabla_x \log p(x) \cdot f(x)^\top + \nabla_x f(x).$$

Stein's Method [Stein, 1972]

- A general theoretical tool for bounding differences between distributions
  - Mostly used for theoretical proof: central limit theorem, concentration inequalities, etc.
- Key idea: Characterizing a distribution p with a Stein operator $\mathcal{A}_p$, such that

$$p = q \quad \Longleftrightarrow \quad \mathbb{E}_{x \sim q}[\mathcal{A}_p f(x)] = 0.$$

  - For continuous distributions with smooth density $p(x)$,

$$\mathcal{A}_p f(x) \stackrel{def}{=} \underbrace{\nabla_x \log p(x)}_{\text{score function}} \cdot f(x)^\top + \nabla_x f(x).$$

    - Score function $\boldsymbol{s}_p(x) = \nabla_x \log p(x) = \frac{\nabla_x p(x)}{p(x)}$, independent of normalization constant Z!

Stein's Method

$p = q$, Stein's Identity : $\mathbb{E}_{x \sim p}[\nabla_x \log p(x) \cdot f(x)^\top + \nabla_x f(x)] = 0$:

Why?

- Use integration by parts, assuming zero boundary conditions.

$$\int p'(x) f(x) + p(x) f'(x) dx = p(x) f(x) \Big|_{-\infty}^{+\infty} = 0.$$

Stein's Method

$p \neq q \quad \Rightarrow \quad \exists$ some f, $\mathbb{E}_{x\sim q}[\mathcal{A}_p f(x)] \neq 0$:

Why?

- We can show (denote $\boldsymbol{s}_p(x) = \nabla_x \log p(x)$):

$$\begin{aligned}\mathbb{E}_{x\sim q}[\mathcal{A}_p f(x)] &= \mathbb{E}_{x\sim q}[\mathcal{A}_p f(x)] - \mathbb{E}_{x\sim q}[\mathcal{A}_q f(x)] \\ &= \mathbb{E}_{x\sim q}[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x)) f(x)^\top]\end{aligned}$$

  - Stein operator: essentially the inner product with the difference of score functions $(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x))$.
  - Unless $\boldsymbol{s}_p(x) \equiv \boldsymbol{s}_q(x)$, we can always find an $f(x)$ to get non-zero.

Stein's Identity

$$\mathbb{E}_{x\sim p}[\nabla_x \log p(x) \cdot f(x)^\top + \nabla_x f(x)] = 0.$$

Stein's identity: an infinite number of identities, indexed by function f

has found lots of applications in machine learning:

- Learning probabilistic models from data
  - Score matching [Hyvärinen, 2005, Lyu, 2009, Sriperumbudur et al., 2013]
  - Spectrum methods [Sedghi and Anandkumar, 2014]
- Variance reduction [Oates et al., 2014, 2016, 2017]
- Feature learning [Janzamin et al., 2014]
- Optimization [Erdogdu, 2015]
- and many more ...

Stein Discrepancy

$$p \neq q \quad \Longrightarrow \quad \exists f, \text{ such that } \mathbb{E}_{x \sim q}[\mathcal{A}_p f(x)] \neq 0$$

- Define (Squared) Stein discrepancy between p and q:

$$\sqrt{\mathbb{S}(q,\ p)} = \max_{f \in \mathcal{F}} \mathbb{E}_{x \sim q}[\mathrm{trace}(\mathcal{A}_p f(x))]$$

- It gives a functional optimization. Computationally difficult for practical use.

Stein Discrepancy

$$p \neq q \quad \Longrightarrow \quad \exists f, \text{ such that } \mathbb{E}_{x\sim q}[\mathcal{A}_p f(x)] \neq 0$$

- Define (Squared) Stein discrepancy between p and q:

$$\sqrt{\mathbb{S}(q,\ p)} = \max_{f\in\mathcal{F}} \mathbb{E}_{x\sim q}[\mathrm{trace}(\mathcal{A}_p f(x))]$$

- It gives a functional optimization. Computationally difficult for practical use.

Kernelized Stein discrepancy (KSD)

- Let $k(x, x')$ be a positive definite kernel, and $\mathcal{H}$ its related reproducing kernel Hilbert space (RKHS). $\mathcal{H}^d = \mathcal{H} \times \cdots \times \mathcal{H}$.
- Kernelized Stein discrepancy: take $\mathcal{F}$ to be the unit ball of RKHS.

$$\sqrt{\mathbb{S}(q,\ p)} = \max_{f \in \mathcal{F}} \mathbb{E}_{x \sim q}[\mathrm{trace}(\mathcal{A}_p f(x))], \qquad f = \{f \in \mathcal{H}^d \colon ||f||_{\mathcal{H}^d} \leq 1\}$$

Kernelized Stein discrepancy (KSD)

- Let $k(x, x')$ be a positive definite kernel, and $\mathcal{H}$ its related reproducing kernel Hilbert space (RKHS). $\mathcal{H}^d = \mathcal{H} \times \cdots \times \mathcal{H}$.
- Kernelized Stein discrepancy: take $\mathcal{F}$ to be the unit ball of RKHS.

$$\sqrt{\mathbb{S}(q,\ p)} = \max_{f \in \mathcal{F}} \mathbb{E}_{x \sim q}[\mathrm{trace}(\mathcal{A}_p f(x))], \qquad f = \{\mathcal{F} \in \mathcal{H}^d \colon ||f||_{\mathcal{H}^d} \leq 1\}$$

then it has a closed form solution

$$\mathbb{S}(q,\ p) = \mathbb{E}_{x,x' \sim q}[\kappa_p(x, x')]$$

where $\kappa_p(x, x') = \mathrm{trace}(\mathcal{A}_p^x \mathcal{A}_p^{x'} k(x, x'))$

$$= \boldsymbol{s}_p(x)^\top k(x, x') \boldsymbol{s}_p(x') + \boldsymbol{s}_p(x)^\top \nabla_{x'} k(x, x') + \nabla_x k(x, x')^\top \boldsymbol{s}_p(x') + \Delta k(x, x')$$

where $\mathcal{A}_p^x$ is Stein operator w.r.t. x.

Kernelized Stein discrepancy (KSD)

- Let $k(x, x')$ be a positive definite kernel, and $\mathcal{H}$ its related reproducing kernel Hilbert space (RKHS). $\mathcal{H}^d = \mathcal{H} \times \cdots \times \mathcal{H}$.
- Kernelized Stein discrepancy: take $\mathcal{F}$ to be the unit ball of RKHS.

$$\sqrt{\mathbb{S}(q,\ p)} = \max_{f \in \mathcal{F}} \mathbb{E}_{x \sim q}[\mathrm{trace}(\mathcal{A}_p f(x))], \qquad f = \{\mathcal{F} \in \mathcal{H}^d : ||f||_{\mathcal{H}^d} \leq 1\}$$

then it has a closed form solution

$$\mathbb{S}(q,\ p) = \mathbb{E}_{x,x' \sim q}[\kappa_p(x, x')] \approx \underbrace{\frac{\mathbf{1}}{\mathbf{n(n-1)}} \sum_{\mathbf{i \neq j}} \kappa_{\mathbf{p}}(\mathbf{x_i}, \mathbf{x_j})}_{\text{empirical estimation}}$$

where $\quad \kappa_p(x, x') = \mathrm{trace}(\mathcal{A}_p^x \mathcal{A}_p^{x'} k(x, x'))$

$$= \boldsymbol{s}_p(x)^\top k(x, x') \boldsymbol{s}_p(x') + \boldsymbol{s}_p(x)^\top \nabla_{x'} k(x, x') + \nabla_x k(x, x')^\top \boldsymbol{s}_p(x') + \Delta k(x, x')$$

where $\mathcal{A}_p^x$ is Stein operator w.r.t. x.

Empirical Estimation and Goodness-of-fit Tests

- Given $\{x_i\} \sim q(x)$, we can get an unbiased estimator of $S(q,\ p)$ by U-statistic:

$$\hat{S}(q,\ p) = \frac{1}{n(n-1)} \sum_{i \neq j} \kappa_p(x_i, x_j).$$

- Asymptotic distribution is well understood:
  - If $p \neq q$, $\hat{S}(q,\ p) = S(q,\ p) + O_p(1/\sqrt{n})$ (asymptotic normal)
  - If $p = q$, $\hat{S}(q,\ p) = O_p(1/n)$. (infinite sum of χ^2 distributions)
- Goodness-of-fit test:
  - Reject $p = q$ if $\hat{S}(q,\ p) > \gamma$.
  - Threshold γ decided using a generalized bootstrap procedure by Arcones and Gine [1992], Huskova and Janssen [1993].

Kernelized Stein Discrepancy (KSD)

- Is KSD a valid discrepancy: $p = q \iff \mathbb{S}(q,\ p) = 0$?

- We can show

$$S(q,\ p) = \mathbb{E}_{x,x' \sim q}[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x))^\top k(x,x')(\boldsymbol{s}_p(x') - \boldsymbol{s}_q(x'))]$$

$$(\text{Recall that } \mathbb{E}_q[\mathcal{A}_p f(x)] = \mathbb{E}_q[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x)) f(x)^\top])$$

- We just need $k(x,x')$ to be integrally strictly positive definite:

$$\int g(x)k(x,x')g(x')dx > 0 \quad \forall g \in L_2 \setminus \{0\}$$

Kernelized Stein Discrepancy (KSD)

- Is KSD a valid discrepancy: $p = q \iff \mathbb{S}(q,\ p) = 0$?
- We can show

$$S(q,\ p) = \mathbb{E}_{x,x' \sim q}[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x))^\top k(x, x')(\boldsymbol{s}_p(x') - \boldsymbol{s}_q(x'))]$$

$$\text{(Recall that } \mathbb{E}_q[\mathcal{A}_p f(x)] = \mathbb{E}_q[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x)) f(x)^\top])$$

- We just need $k(x, x')$ to be integrally strictly positive definite:

$$\int g(x) k(x, x') g(x') dx > 0 \quad \forall g \in L_2 \setminus \{0\}$$

Kernelized Stein Discrepancy (KSD)

- Is KSD a valid discrepancy: $p = q \iff \mathbb{S}(q,\ p) = 0$?
- We can show

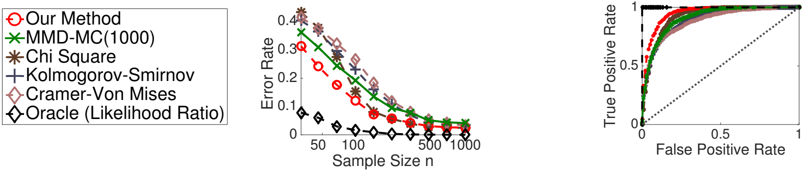
$$S(q,\ p) = \mathbb{E}_{x,x'\sim q}[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x))^\top k(x,x')(\boldsymbol{s}_p(x') - \boldsymbol{s}_q(x'))]$$

(Recall that $\mathbb{E}_q[\mathcal{A}_p f(x)] = \mathbb{E}_q[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x)) f(x)^\top])$

- We just need $k(x,x')$ to be <u>integrally strictly positive definite</u>:

$$\int g(x)k(x,x')g(x')dx > 0 \quad \forall g \in L_2 \setminus \{0\}$$

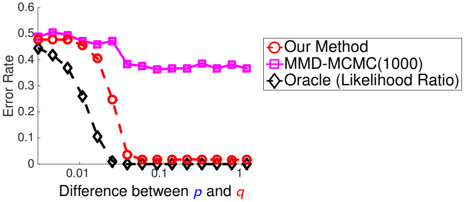
- 1D Gaussian mixture model (GMM)
  - Simulate samples either from the true, or the perturbed model with equal probabilities
  - Use GOF tests to tell if the sample is drawn from the true model (significance $\alpha = 0.05$).



(Fixed $n = 100$)

Gaussian-Bernoulli Restricted Boltzmann Machine

- Gaussian visible nodes + binary hidden nodes.
- Computationally intractable to draw exact sample, or calculate the normalization constant (and likelihood).



(50 visible nodes; 10 hidden nodes)

Connection with Other Discrepancy Measures

Maximum mean discrepancy (MMD)

- Maximum mean discrepancy (MMD):

$$\mathbb{M}(q,\ p) = \max_{f \in \mathcal{H}} \{\mathbb{E}_p f - \mathbb{E}_q f \qquad s.t. \quad ||f||_{\mathcal{H}} \leq 1\}.$$

$\mathcal{H}$ is the RKHS related to $k(x, x')$.

- KSD can be treated as a MMD using the "Steinalized" kernel $\kappa_p(x, x') = \mathrm{trace}(\mathcal{A}_p^x \mathcal{A}_p^{x'} k(x, x'))$, which depends on p (KSD is asymmetric):

$$\mathbb{S}(q,\ p) = \max_{f \in \mathcal{H}_p} \{\mathbb{E}_p f - \mathbb{E}_q f \qquad s.t. \quad ||f||_{\mathcal{H}_p} \leq 1\}$$

$\mathcal{H}_p$ is the RKHS related to $k_p(x, x')$.

Connection with Other Discrepancy Measures

Fisher divergence

- Fisher divergence: $\mathbb{F}(q,\ p) = \mathbb{E}_{x \sim q}[||\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x)||_2^2]$.
- Used as a learning objective in score matching.
- KSD is a smoothed version of Fisher divergence; we can show

$$\mathbb{S}(q,\ p) = \mathbb{E}_{x \sim q}[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x))^\top k(x, x')(\boldsymbol{s}_p(x') - \boldsymbol{s}_q(x'))].$$

Connection with Other Discrepancy Measures

Fisher divergence

- Fisher divergence: $\mathbb{F}(q,\ p) = \mathbb{E}_{x\sim q}[||\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x)||_2^2]$.
- Used as a learning objective in score matching.
- KSD is a smoothed version of Fisher divergence; we can show

$$\mathbb{S}(q,\ p) = \mathbb{E}_{x\sim q}[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x))^\top k(x,x')(\boldsymbol{s}_p(x') - \boldsymbol{s}_q(x'))].$$

KL Divergence

- Fisher divergence = derivative of KL when variables are perturbed by i.i.d. Gaussian.
- KSD = derivative of KL when variables are perturbed by smooth functions in RKHS (see Liu & Wang NIPS 2016, where a variational inference method based on it).

Related Work

- Chwialkowski et al. [2016]: Independent work on quite the same idea (this ICML, the talk before us).
- Oates et al. [2014, 2016, 2017]: Combined Stein's identity with RKHS; used for deriving a super-efficient variance reduction method.
- Gorham and Mackey [2015]: Derived a different (non-kernel) computable Stein discrepancy by enforcing smoothness constraints on a finite number of points, solved by linear programming.

Related Work

- Chwialkowski et al. [2016]: Independent work on quite the same idea (this ICML, the talk before us).
- Oates et al. [2014, 2016, 2017]: Combined Stein's identity with RKHS; used for deriving a super-efficient variance reduction method.
- Gorham and Mackey [2015]: Derived a different (non-kernel) computable Stein discrepancy by enforcing smoothness constraints on a finite number of points, solved by linear programming.

Related Work

- Chwialkowski et al. [2016]: Independent work on quite the same idea (this ICML, the talk before us).
- Oates et al. [2014, 2016, 2017]: Combined Stein's identity with RKHS; used for deriving a super-efficient variance reduction method.
- Gorham and Mackey [2015]: Derived a different (non-kernel) computable Stein discrepancy by enforcing smoothness constraints on a finite number of points, solved by linear programming.

Conclusion

- Defined and studied kernelized Stein discrepancy (KSD)
- Derived a goodness-of-fit test that works for distributions with intractable normalization constants
- Directions:
  - More understandings and applications

Conclusion

- Defined and studied kernelized Stein discrepancy (KSD)
- Derived a goodness-of-fit test that works for distributions with intractable normalization constants
- Directions:
  - More understandings and applications

Thank You

References I

M. A. Arcones and E. Gine. On the bootstrap of U and V statistics. The Annals of Statistics, pages 655–674, 1992.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In ICML, 2016.

M. A. Erdogdu. Newton-stein method: An optimization method for glms via stein's lemma. arXiv preprint arXiv:1511.08895, 2015.

J. Gorham and L. Mackey. Measuring sample quality with stein's method. In NIPS, pages 226–234, 2015.

M. Huskova and P. Janssen. Consistency of the generalized bootstrap for degenerate U-statistics. The Annals of Statistics, pages 1811–1823, 1993.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. In Journal of Machine Learning Research, pages 695–709, 2005.

M. Janzamin, H. Sedghi, and A. Anandkumar. Score function features for discriminative learning: Matrix and tensor framework. arXiv preprint arXiv:1412.2863, 2014.

S. Lyu. Interpretation and generalization of score matching. In UAI, pages 359–366, 2009.

C. J. Oates, M. Girolami, and N. Chopin. Control functionals for monte carlo integration. arXiv preprint arXiv:1410.2392, 2014.

C. J. Oates, J. Cockayne, F.-X. Briol, and M. Girolami. Convergence rates for a class of estimators based on stein's identity. arXiv preprint arXiv:1603.03220, 2016.

C. J. Oates, M. Girolami, and N. Chopin. Control functionals for monte carlo integration. Journal of the Royal Statistical Society, Series B, 2017.

H. Sedghi and A. Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. arXiv preprint arXiv:1412.3046, 2014.

B. Sriperumbudur, K. Fukumizu, R. Kumar, A. Gretton, and A. Hyvärinen. Density estimation in infinite dimensional exponential families. arXiv preprint arXiv:1312.3516, 2013.

C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory, pages 583–602, 1972.