High-dimensional test for normality

Jérémie Kellner
Ph.D Student

University Lille I - MODAL project-team Inria

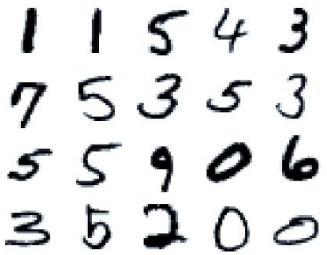
joint work with Alain Celisse

Rennes - June 5th, 2014

Framework

Input space $\mathcal{X}$ of any kind:

- Scalars or vectors,
- Structured objects (strings, graphs, trees, ...),
- Functional space, ...



$\mathcal{X}$: handwritten digits

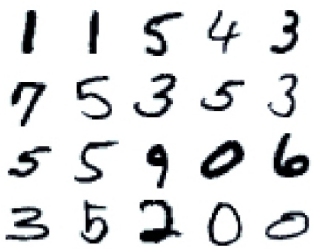
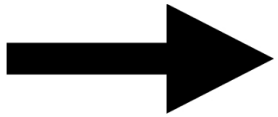
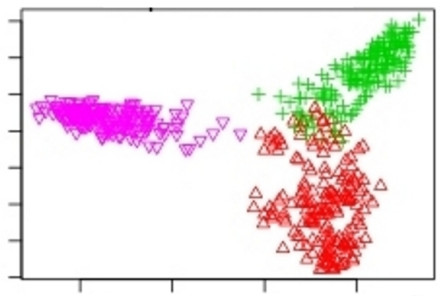
Working in kernel space

- $X_1, \ldots, X_n \in \mathcal{X}$ i.i.d.
- Positive semi-definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
- Mappings $Y_i = k(X_i, .) \in H(k)$

Definition (RKHS)

- $H(k) = \overline{\mathrm{Span}\{k(x, .) \mid x \in \mathcal{X}\}}$
- Reproducing property:

$$\forall x, y \in \mathcal{X}, \ < k(x, .), k(y, .) >_{H(k)} = k(x, y)$$

X_i    $Y_i = k(X_i, .)$

Gaussian process in RKHS

Gaussian assumption in high-dimensional/kernel spaces

- Mean equality test in a high-dimensional space (Srivastava et al., 2013)
- Supervised/unsupervised classification using Gaussian mixtures in kernel space (Bouveyron et al., 2012)

Gaussian process in RKHS

Gaussian assumption in high-dimensional/kernel spaces

- Mean equality test in a high-dimensional space (Srivastava et al., 2013)
- Supervised/unsupervised classification using Gaussian mixtures in kernel space (Bouveyron et al., 2012)

Gaussian process

$$Z \sim \mathcal{GP}(\mu, \Sigma) \quad \textbf{iff} \quad \forall h \in H(k), \ < Z, h > \sim \mathcal{N}(< \mu, h >, < \Sigma h, h >)$$

Gaussian process in RKHS

Gaussian assumption in high-dimensional/kernel spaces

- Mean equality test in a high-dimensional space (Srivastava et al., 2013)
- Supervised/unsupervised classification using Gaussian mixtures in kernel space (Bouveyron et al., 2012)

Gaussian process

$$Z \sim \mathcal{GP}(\mu, \Sigma) \quad \textbf{iff} \quad \forall h \in H(k), \ < Z, h > \sim \mathcal{N}(< \mu, h >, < \Sigma h, h >)$$

Goal

Test $\mathcal{H}_0 : P = P_0$ vs $\mathcal{H}_A : P \neq P_0$, where $P_0 = \mathcal{GP}(\mu, \Sigma)$

Outline

Distinguishing distributions with MMD

MMD (Gretton et al., 2007)

Y, Z two r.v. in any set $\mathcal{X}$.

$$MMD(Y,Z) = \sup_{f \in H(k), ||f|| \leq 1} |\mathbb{E}_Y f(Y) - \mathbb{E}_Z f(Z)|$$

- **Advantage:** *MMD* can be computed as a distance between two elements of $H(k)$ (easy calculation),
- **Problem:** *MMD* is a metric on distributions only for some k (*characteristic kernels*).

Consider Laplace transforms of P and P_0 on $H(k)$

$$\mathcal{L}_P(f) \triangleq \mathbb{E}_{Y\sim P} e^{<Y,f>_{H(k)}} \quad , \quad \mathcal{L}_{P_0}(f) \triangleq \mathbb{E}_{Z\sim P_0} e^{<Z,f>_{H(k)}}$$

Consider Laplace transforms of P and P_0 on $H(k)$

$$\mathcal{L}_P(f) \triangleq \mathbb{E}_{Y\sim P} e^{<Y,f>_{H(k)}} \quad , \quad \mathcal{L}_{P_0}(f) \triangleq \mathbb{E}_{Z\sim P_0} e^{<Z,f>_{H(k)}}$$

Compare $\mathcal{L}_P$ with $\mathcal{L}_{P_0}$

$$\Delta(P, P_0) \triangleq \sup_{||f||\leq 1} |\mathcal{L}_P(f) - \mathcal{L}_{P_0}(f)|$$

Consider Laplace transforms of P and P_0 on $H(k)$

$$\mathcal{L}_P(f) \stackrel{\Delta}{=} \mathbb{E}_{Y\sim P} e^{<Y,f>_{H(k)}} \quad , \quad \mathcal{L}_{P_0}(f) \stackrel{\Delta}{=} \mathbb{E}_{Z\sim P_0} e^{<Z,f>_{H(k)}}$$

Compare $\mathcal{L}_P$ with $\mathcal{L}_{P_0}$

$$\Delta(P, P_0) \stackrel{\Delta}{=} \sup_{||f||\leq 1} |\mathcal{L}_P(f) - \mathcal{L}_{P_0}(f)|$$

We get the desired property

$$\Delta(P, P_0) = 0 \implies P = P_0$$

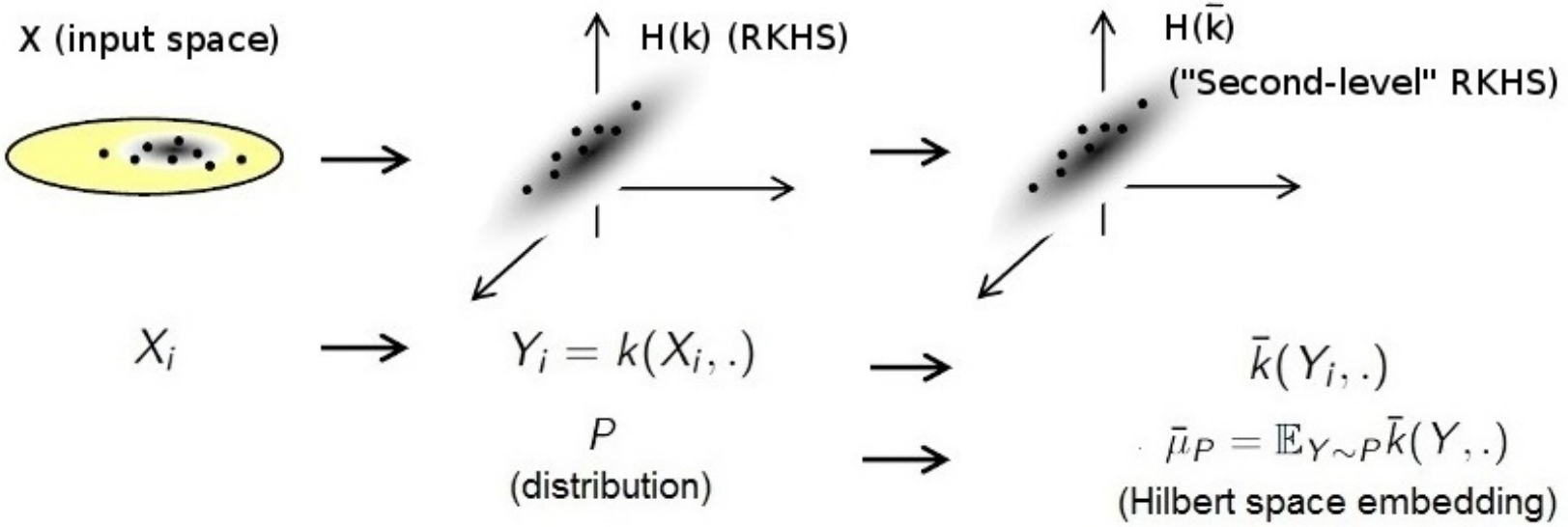
without requiring that k is characteristic.

Introducing a second RKHS

Get a computable expression for

$$\Delta(P, P_0) = \sup_{||f|| \leq 1} \left| \mathbb{E}_Y \bar{k}(Y, f) - \mathbb{E}_Z \bar{k}(Z, f) \right|$$

via kernel $\bar{k} = \exp(< ., . >_{H(k)})$



$$\Delta(P, P_0) = \sup_{||f|| \leq 1} \left| \mathbb{E}_Y \bar{k}(Y, f) - \mathbb{E}_Z \bar{k}(Z, f) \right|$$

$$\Delta(P, P_0) = \sup_{||f|| \leq 1} \left| \mathbb{E}_Y \bar{k}(Y, f) - \mathbb{E}_Z \bar{k}(Z, f) \right|$$
$$= \sup_{||f|| \leq 1} |\mathbb{E}_P < \bar{k}(Y, .), \bar{k}(f, .) > - \mathbb{E}_{P_0} < \bar{k}(Z, .), \bar{k}(f, .) > |$$

(from reproducing property)

$$\begin{aligned}
\Delta(P, P_0) = & \sup_{||f|| \leq 1} \left| \mathbb{E}_Y \bar{k}(Y, f) - \mathbb{E}_Z \bar{k}(Z, f) \right| \\
= & \sup_{||f|| \leq 1} \left| \mathbb{E}_P < \bar{k}(Y, .), \bar{k}(f, .) > - \mathbb{E}_{P_0} < \bar{k}(Z, .), \bar{k}(f, .) > \right| \\
& \qquad \text{(from reproducing property)} \\
= & \sup_{||f|| \leq 1} \left| < \bar{\mu}_P - \bar{\mu}_{P_0}, \bar{k}(f, .) >_{H(\bar{k})} \right|
\end{aligned}$$

$$\begin{aligned}
\Delta(P, P_0) = & \sup_{||f|| \leq 1} \left| \mathbb{E}_Y \bar{k}(Y, f) - \mathbb{E}_Z \bar{k}(Z, f) \right| \\
= & \sup_{||f|| \leq 1} |\mathbb{E}_P < \bar{k}(Y, .), \bar{k}(f, .) > - \mathbb{E}_{P_0} < \bar{k}(Z, .), \bar{k}(f, .) > | \\
& \qquad \text{(from reproducing property)} \\
= & \sup_{||f|| \leq 1} | < \bar{\mu}_P - \bar{\mu}_{P_0}, \bar{k}(f, .) >_{H(\bar{k})} | \\
\leq & \; e^{1/2} \; ||\bar{\mu}_P - \bar{\mu}_{P_0}||_{H(\bar{k})} \\
& \qquad \text{(from Cauchy-Schwarz)}
\end{aligned}$$

$$\begin{aligned}
\Delta(P, P_0) = & \sup_{||f|| \leq 1} \left| \mathbb{E}_Y \bar{k}(Y, f) - \mathbb{E}_Z \bar{k}(Z, f) \right| \\
= & \sup_{||f|| \leq 1} |\mathbb{E}_P < \bar{k}(Y, .), \bar{k}(f, .) > - \mathbb{E}_{P_0} < \bar{k}(Z, .), \bar{k}(f, .) > | \\
& \qquad \text{(from reproducing property)} \\
= & \sup_{||f|| \leq 1} | < \bar{\mu}_P - \bar{\mu}_{P_0}, \bar{k}(f, .) >_{H(\bar{k})} | \\
\leq & \; e^{1/2} \; ||\bar{\mu}_P - \bar{\mu}_{P_0}||_{H(\bar{k})} \\
& \qquad \text{(from Cauchy-Schwarz)}
\end{aligned}$$

Definition (Laplace-MMD)

Assume $\max(\mathbb{E}_P e^{||Y||^2/2}, \mathbb{E}_{P_0} e^{||Z||^2/2}) < +\infty$.

$$L \stackrel{\Delta}{=} ||\bar{\mu}_P - \bar{\mu}_{P_0}|| = 0 \Leftrightarrow P = P_0$$

$\Rightarrow$ L is an easy-to-handle quantity:

- $\bar{\mu}_P$ estimated by $\bar{\mu}_{\hat{P}}$ (sample mean)
- Expand the (squared) norm

L-MMD test

Gram matrix: $K = [k(X_i, X_j)]_{i,j}$

Proposition (K., 2013)

Assume $P_0 = \mathcal{GP}(0, \Sigma)$ and $\rho(\Sigma) < 1$. Then,

$$n\hat{L}^2 = \frac{1}{n-1}\sum_{i \neq j}^{n} e^{K_{i,j}} - 2\sum_{i=1}^{n} e^{[K^2]_{i,i}/(2n)} + n\left[\det(I - n^{-2}K^2)\right]^{-1/2}$$

is an unbiaised estimator of nL^2.

L-MMD test

Gram matrix: $K = [k(X_i, X_j)]_{i,j}$

Proposition (K., 2013)

Assume $P_0 = \mathcal{GP}(0, \Sigma)$ and $\rho(\Sigma) < 1$. Then,

$$n\hat{L}^2 = \frac{1}{n-1}\sum_{i \neq j}^{n} e^{K_{i,j}} - 2\sum_{i=1}^{n} e^{[K^2]_{i,i}/(2n)} + n\left[\det(I - n^{-2}K^2)\right]^{-1/2}$$

is an unbiaised estimator of nL^2.

Rejection region

- Generate $n\hat{L}^2_{(1)} \leq \ldots \leq n\hat{L}^2_{(B)}$ under $\mathcal{H}_0$
- Set $\hat{q}_{\alpha,n} := n\hat{L}^2_{(t)}$ where $t = t(\alpha)$
- Reject $\mathcal{H}_0$ if $n\hat{L}^2 \geq \hat{q}_{\alpha,n}$, accept otherwise.

Outline

Type-II error: theoretical bound

Theorem (K., 2014): If $||Y|| \leq M$ P-a.s. Then for $n > \frac{q_{\alpha,n}+m_P^{(2)}}{L^2}$

$$\mathbb{P}_{\mathcal{H}_A}(n\hat{L}^2 \leq \hat{q}_{\alpha,n}) \leq \left[1 + o_B(1/\sqrt{B})\right] \exp\left(-\frac{n\left\{L - \sqrt{\frac{\square}{n-1}}\right\}^2}{\Delta}\right)$$

where

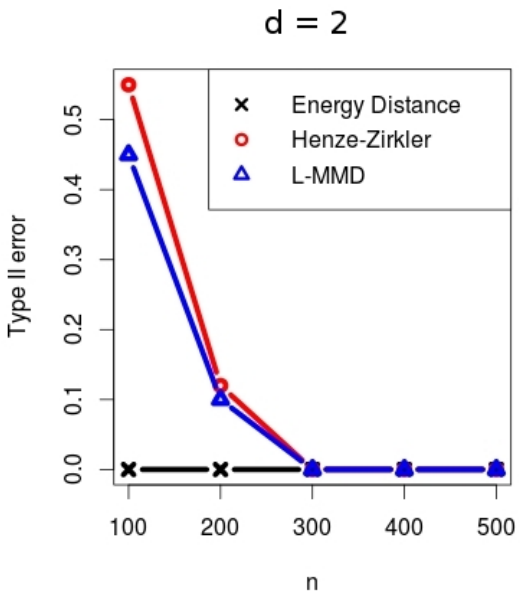
$$\square = q_{\alpha,n} + m_P^{(2)}$$

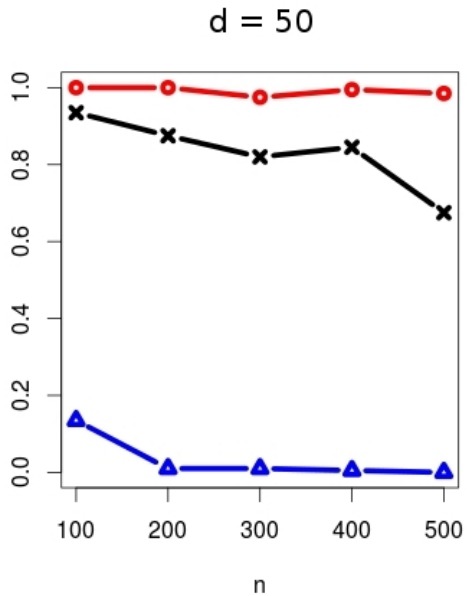
$$\Delta = 2m_P^{(2)} + \frac{16}{3}\sqrt{m_P^{(2)}}L^2 \exp(M^2/2) + o_n(1)$$

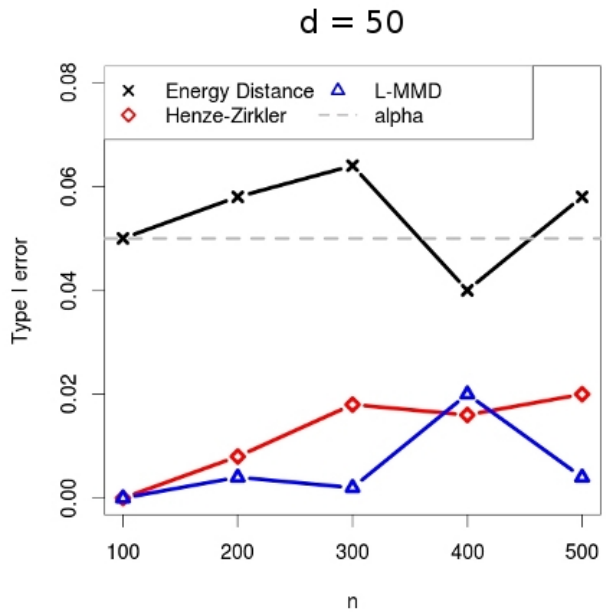
$$m_P^{(2)} = \mathbb{E}_{Y\sim P}||\bar{k}(Y,\cdot) - \bar{\mu}_P||^2_{H(\bar{k})} = \mathbb{E}||\bar{k}(Y,\cdot) - \mathbb{E}\left[\bar{k}(Y,\cdot)\right]||^2_{H(\bar{k})}$$

Synthetic data (finite d):

- $\mathcal{X} = \mathbb{R}^d$, $k =< ., . >_{\mathbb{R}^d}$: L-MMD used as a multivariate normality test
- Common multivariate normality tests lose power when d large
  1. Henze-Zirkler (characteristic functions, L_2 distance)
  2. Energy distance (pairwise distance)
- Alternative: mixture of two Gaussians $\mathcal{N}(\mu_1, \Sigma)$ and $\mathcal{N}(\mu_2, \Sigma)$
- Two cases: low dimension ($d = 2$), larger dimension ($d = 50$)



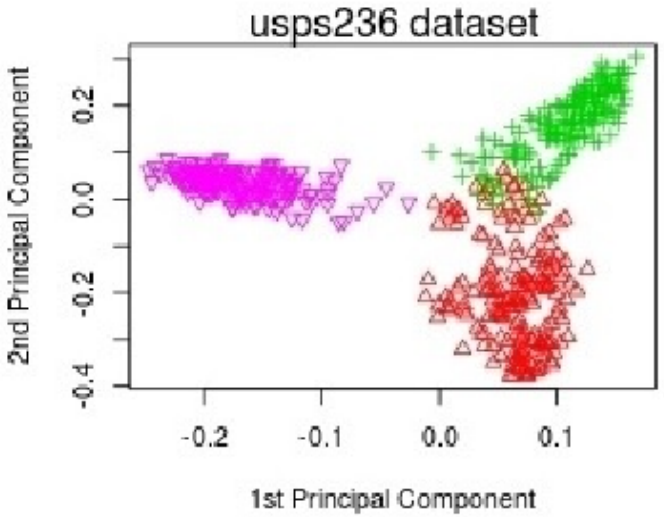


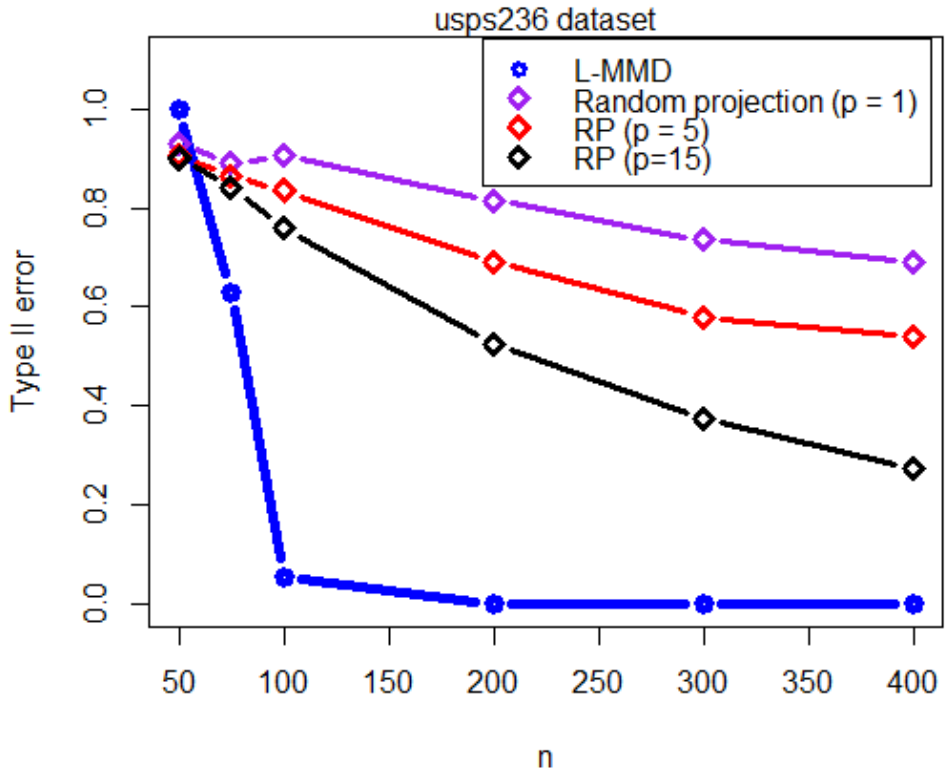


Real data ($d = +\infty$):

- USPS236 dataset $\Rightarrow$ input space $\mathcal{X} = \mathbb{R}^{64}$
- Gaussian kernel $k(x, y) = \exp(-(2\sigma^2)^{-1}||x - y||^2)$
- Compare L-MMD with Random Projection method

  $\Longrightarrow$ Kolmogorov-Smirnov (univariate) test on p random projections





Conclusion

Summary:

- High-dimensional test for normality
- Bypassed characteristic assumption
- Mild sensitivity to high-dimensionality

Further works:

- In practice, μ and Σ unknown
  - $\longrightarrow$ How does parameters estimations affect Type-I/II errors?
  - $\longrightarrow$ Type-I adjustement method within this framework?
- Extension to two-sample homogeneity test

Conclusion

Summary:

- High-dimensional test for normality
- Bypassed characteristic assumption
- Mild sensitivity to high-dimensionality

Further works:

- In practice, μ and Σ unknown
  - $\longrightarrow$ How does parameters estimations affect Type-I/II errors?
  - $\longrightarrow$ Type-I adjustement method within this framework?
- Extension to two-sample homogeneity test

Merci pour votre attention.