Predictive Density Estimation : recent results

Éric Marchand

(Banff International Research Station, April 9, 2019)

(Collaborators : Tatsuya Kubokawa, Bill Strawderman, Dominique Fourdrinier, Aziz L'Moudden)

OUTLINE

1. INTRODUCTION
2. LOSS and RISK
3. PLUG-IN, DUALITY, and EXTENSIONS
4. APPLICATIONS
5. IMPROVEMENTS BY SCALE EXPANSION
6. SPHERICALLY SYMMETRIC DISTRIBUTIONS with UNKNOWN LOCATION and SCALE
7. SUMMARY AND CONCLUDING REMARKS

1. **INTRODUCTION**

— Suppose $X, Y \in \mathbb{R}^d$, continuous, with $X \sim p_\theta$, $Y \sim q_\theta$. X is observed, Y unobserved or missing.

— We seek a predictive density estimator $\hat{q}(\cdot; X)$ for q_θ, and assess its performance via a *distance* ρ and corresponding loss $L(\theta, \hat{q}) = \rho(q_\theta, \hat{q})\,, \theta \in \Theta$.

— Bayesian Approach $\Rightarrow$ prior π and posterior density $\pi(\cdot|x)$, so that a natural estimator for q_θ is (when X and Y are conditionally independent on θ)

$$q_\pi(y|x) = \int_\Theta q_\theta(y)\,\pi(\theta|x)d\theta$$

i.e., a posterior distribution mixture of q_θ.

— The estimator $q_\pi(y|x)$ is attractive from a Bayesian perspective (e.g., Jeffreys, 1939). Also formal Bayes rule for both integrated L_2 and Kullback-Leibler (KL) losses.

2. **LOSS and RISK**

Several loss functions are at our disposal :

(I) Kullback-Leibler.

$$L_{KL}(\theta, \widehat{q}) = \int_{\mathbb{R}^d} q_\theta(y) \log \frac{q_\theta(y)}{\widehat{q}(y)} \, dy$$

(II) $\alpha-$divergence (Csiszár, 1967).

$$L_{h_\alpha}(\theta, \widehat{q}) = \int_{\mathbb{R}^d} h_\alpha \left(\frac{\widehat{q}(y)}{q_\theta(y)} \right) q_\theta(y) \, dy \, ,$$

with $h_\alpha(z) = \begin{cases} \frac{4}{1-\alpha^2}(1 - z^{\frac{(1+\alpha)}{2}}) & |\alpha| \leq 1 \\ z \log(z) & \alpha = 1 \\ -\log(z) & \alpha = -1. \end{cases}$

Hellinger ($\alpha = 0$), Kullback-Leibler ($\alpha = -1$), Reverse Kullback-Leibler ($\alpha = 1$)

(III) Integrated L^s, namely L^1 and L^2.

$$L^s(\theta, \widehat{q}) = \int_{\mathbb{R}^d} |\widehat{q}(y) - q_\theta(y)|^s \, dy \, .$$

Frequentist risk can be evaluated by

$$R(\theta, \widehat{q}) = \mathbb{E}_\theta \, L(\theta, q(\cdot; X))$$

.

3. PLUG-IN, DUALITY, and EXTENSIONS

— Plug-in predictive density estimator of q_θ is $q_{\widehat{\theta}}$, where $\widehat{\theta}(X)$ is a point estimator of θ. Includes $\widehat{q}_{\text{mle}} = q_{\widehat{\theta}_{\text{mle}}}$.

— Loss incurred is $L(\theta, q_{\widehat{\theta}})$, i.e. a point estimation loss incurred by $\widehat{\theta}$ in estimating θ.

— **Duality.** The efficiency of the plug-in $q_{\widehat{\theta}}$ for estimating q_θ is dual to the efficiency of $\widehat{\theta}$ for estimating θ.

— **Example.** $q_\theta \sim Gamma(a, \theta)$. Dual loss for plug-in's is :

$$L(\theta, q_{\widehat{\theta}}) = a\left(\frac{\theta}{\widehat{\theta}} - \log\frac{\theta}{\widehat{\theta}} - 1\right)$$

— The above idea extends to the comparison of predictive densities of the form $f_{\widehat{\theta}}$ with dual loss given by $L(\theta, f_{\widehat{\theta}})$.

— Here are several dominance results derived with the above.

4. **APPLICATIONS**

(a) $X \sim N_d(\theta, \sigma_X^2 I_d)$, $Y \sim N_d(\theta, \sigma_Y^2 I_d)$. Consider

$$f_{\widehat{\theta}} \sim N_d(\widehat{\theta}(X), (\sigma_X^2 + \sigma_Y^2) I_d)\,.$$

Under KL loss, $\widehat{\theta}_0(X) = X$ yields the MRE predictive density (also minimax). Dual loss is $L(\theta, f_{\widehat{\theta}}) \propto \|\widehat{\theta} - \theta\|^2$. So, for $d \geq 3$, dominating estimators of $\widehat{\theta}_0$, yield improvements on $\widehat{q}_{\text{mre}}$. Inadmissibility of $\widehat{q}_{\text{mre}}$ is due to Komaki (2001).

(b) Same as (a) for reverse Kullback-Leibler, but with $\widehat{q}_{\text{mre}} \sim N_d(X, \sigma_Y^2 I_d)$.

(c) Model as in **(a)** , $\alpha-$divergence loss. Consider the subclass

$$f_{\widehat{\theta}} \sim N_d(\widehat{\theta}(X), (\frac{(1-\alpha)}{2}\sigma_X^2 + \sigma_Y^2) I_d)\,,$$

which includes $\widehat{q}_{\text{mre}}$ (also minimax) for $\widehat{\theta}(X) = X$. Here, reflected normal loss L_γ is dual, with $L_\gamma(\theta, \widehat{\theta}) = 1 - e^{-\frac{\|\widehat{\theta}-\theta\|^2}{2\gamma}}$, for some γ function of $\sigma_X^2, \sigma_Y^2, \alpha$. For $d \geq 3$, dominating estimators of X (KMS, 2015), yield dominating predictive densities $f_{\widehat{\theta}}$ of $f_{\widehat{\theta}_0}$.

Lemma 1. *(KMS 2015) Let* $X \sim N_d(\theta, \sigma_X^2 I_d)$ *with known* σ_X^2*.* $\hat{\theta}(X)$ *dominates* X *under* L_γ *whenever* $\hat{\theta}(Z)$ *dominates* Z *for* $Z \sim N_d(\theta, \frac{\gamma\sigma_X^2}{\gamma+1} I_d)$ *under loss* $\|\hat{\theta} - \theta\|^2$*.*

(d) Same as (c) for L^2 loss. Similar results, but with $\hat{q}_{\text{mre}} \sim N_d(X, (\sigma_X^2 + \sigma_Y^2) I_d)$.

(e) Consider $Y = (Y_1, \ldots, Y_d)' \sim q(\|y - \theta\|^2)$ with unimodal q and L^1 loss. Consider plug-in predictive densities $q(\|y - \hat{\theta}(X)\|^2)$.

Lemma 2. *(KMS, 2017; Dasgupta & Lahiri, 2012) Dual loss is given by*

$$\int_{\mathbb{R}^d} |q(\|y-\hat{\theta}\|^2) - q(\|y-\theta\|^2)|\, dy = 4F(\frac{\|\hat{\theta} - \theta\|}{2}) - 2\,,$$

where $F(t) = P(Y_1 \le t)$*, is the cdf of* Y_1*.*

Via the dual point estimation problem, KMS (2017) obtain for $d \geq 4$ dominating predictive densities $q(\|y - \hat{\theta}(X)\|^2)$ of $\hat{q}_{\text{mle}} \sim q(\|y - X\|^2)$ using Stein estimation techniques for losses which are concave in $\|\hat{\theta} - \theta\|^2$, and with further developments for scale mixtures of normals.

TECHNICAL DETAILS : A SKETCH

A. $X \sim f(\|x-\theta\|^2)$ scale mixture of normals (SMN) ; loss is $\rho(\|\delta-\theta\|^2)$ with ρ' completely monotone (CM)

$$(-1)^n \rho^{(n+1)} \geq 0 \text{ for } n = 0, 1, \ldots$$

B. Concave ρ, $\rho(b) - \rho(a) \leq \rho'(a)(b-a)$.

$\rho(\|\delta-\theta\|^2) - \rho(\|X-\theta\|^2) \leq \rho'(\|X-\theta\|^2)\,\{\|\delta-\theta\|^2 - \|X-\theta\|^2$

$$R_\rho(\theta, \delta) - R_\rho(\theta, X) \leq 0$$

$$\Longleftarrow \mathbb{E}_{f^*}\{\|\delta(X)-\theta\|^2) \; - \; \|X-\theta\|^2\} \leq 0$$

with

$$X \sim f^*(\|x-\theta\|^2) \propto f(\|x-\theta\|^2)\rho'(\|x-\theta\|^2).$$

C.

$$f \text{ is a } SMN \text{ density} \iff f \text{ is CM}$$

$$f \text{ is CM and } \rho' CM \implies f\rho' CM$$

$$\implies f^* \text{ is a } SMN \text{ density}$$

D.

(1) Strawderman (1974) : Conditions for which $\delta(X)$ dominates X under $\|\delta - \theta\|^2$ for $X \sim f^*$ (SMN)

(2) Implies conditions for which $\delta(X)$ dominates X under $\rho(\|\delta - \theta\|^2)$ for $X \sim f$

(3) Implies conditions for which the density $q(\|y - \delta\|^2)$ dominates the plug-in $q(\|y - x|^2)$ for estimating $q(\|y - \theta\|^2)$ for L_1 loss and based on $X \sim p(\|y - \theta\|^2)$ with p, q SMN.

(f) MRE ESTIMATORS

— Location models $X \sim p(x-\theta)$, $Y \sim q(y-\theta)$, indep., $x, y, \theta \in \mathbb{R}^d$.

— Choice of uniform prior $\pi(\theta) = 1$ yields a Bayes estimator which is MRE and Minimax.

— **LEMMA.** For KL and L^2 losses,

$$\begin{aligned} q_{\text{mre}}(y; x) &= \int_{\mathbb{R}^d} q(y-\theta)\, p(x-\theta)\, d\theta \\ &= (q * g)(y-x)\,, \end{aligned}$$

where $g(t) = p(-t)$, $q * g$ is convolution.

— To search for improved predictive density estimators, we consider

$$f_{\hat{\theta}} \sim (q * p)(y - \hat{\theta}(X))\,.$$

For L^2 loss, dual loss is

$$\int_{\mathbb{R}^d} (p(t-\hat{\theta}) - q(t-\theta))^2\, dt =$$

$$q * q(0) + p * p(0) - 2f * q(\theta - \hat{\theta})\,.$$

— KMS (2015) provide improvements for scale mixture of normals and $d \geq 3$ by analyzing this loss.

5. IMPROVEMENTS by SCALE EXPANSION

— $X \sim p(x-\theta)$, $Y \sim q(y-\theta)$, $x, y, \theta \in \mathbb{R}^d$. Plug-in density is $q(y-\hat{\theta})$ where $\hat{\theta}$ is an estimator of θ.

— For Kullback-Leibler or L_2 loss, Bayes predictive density is $\hat{q}_\pi(y|x) = \int_{\mathbb{R}^d} q(y-\theta)\, \pi(\theta|x)\, d\nu(\theta)$

— Useful insight is : (illustrated for $d=1$) For densities $Y \sim q(t-\theta)$ vs. $Y \sim \hat{q}_\pi(\cdot|x)$:

$$E_\theta(Y) = E_0(Y) + \theta\,,\, Var_\theta(Y) = \sigma_Y^2$$

$$E_{\hat{q}_\pi}(Y) = E_0(Y) + E(\theta|x)\,,$$

$$Var_{\hat{q}_\pi}(Y) = \sigma_Y^2 + Var(\theta|x)\,.$$

(assuming expectations and variances exist)

— Hence, for such losses, Bayes density estimators always inflate the variance (unless $\theta|x$ is degenerate). And plug-in densities are **not** Bayes. Similar analysis for $d > 1$.

— Deficiency of plug-in estimators not a new theme and also depend on loss.
— For reverse Kullback-Leibler loss and exponential families, Bayes estimators are **always** plug-in estimators !(Yanagimoto, Ohnishi, 2009)

EXAMPLE Consider $X \sim N_d(\theta, \sigma_X^2 I_d)$, $Y \sim N_d(\theta, \sigma_Y^2 I_p)$, KL loss.

THEOREM. Let $\widehat{\theta}(X)$ be an estimator of $\theta \in C$, with risk $R(\theta, \widehat{\theta}) = E(\|\widehat{\theta}(X) - \theta\|^2)$ and $\underline{R} = \inf_{\theta \in C} R(\theta, \widehat{\theta}) > 0$. Let $\widehat{q}_c \sim N_d(\widehat{\theta}(X), c\sigma_Y^2 I_d)$. Then the density $\widehat{q}_c$ dominates the plug-in density $\widehat{q}_1$ under KL loss if $1 < c \leq (1 + \frac{\underline{R}}{d\sigma_Y^2})$, and iff $1 < c \leq c_0(1 + \frac{\underline{R}}{d\sigma_Y^2})$, with $c_0(m)$ the root in c of $(1 - \frac{1}{c})m - \log(c)$ on (m, ∞).

Remarks

— General $\widehat{\theta}$, d. $\widehat{\theta}(X)$ can be proper Bayes, Generalized Bayes, MLE, Shrinkage or Stein estimator, etc.

— C can be $\mathbb{R}^d$, or a subset (i.e., restricted parameter space) of $\mathbb{R}^d$.

— Inflation of variance $\Leftrightarrow$ performance of $\widehat{\theta}(X)$ for estimating θ.

— $c_0(m) \geq m^2$ for all $m > 1$.

— Dominating predictive densities are not Bayesian, but they do extend to scale mixtures of normals

$$\int_1^{c_0} \widehat{q}_c \, dF(c) \, .$$

EXAMPLE. Similar result with Aziz LMoudden for $\alpha-$divergence, $-1 < \alpha < 1$. Applies to a large class of plug-in densities (e.g., James-Stein and other shrinkage estimators).

EXAMPLE. (Gamma model ; LMoudden et al. 2017.)

$X \sim Ga(\alpha_1, \theta), Y \sim Ga(\alpha_2, \theta)$, α_1, α_2 known .

Consider any non-degenerate estimator $\widehat{\theta}(X)$ of θ and the subclass of predictive densities $\widehat{q}_c \sim Ga(\frac{\alpha_2}{c}, c\widehat{\theta}(X)), c \geq 1$. A rationale for the choice lies in the fact that for all x :

$$\mathbb{E}_{\widehat{q}_c}(Y) = \alpha_2 \widehat{\theta}(x) \, , \, Var_{\widehat{q}_c}(Y) = c\alpha_2 (\widehat{\theta}(x))^2 \, ,$$

which expands the variance as c increases. A key finding is that $\widehat{q}_c$ dominates the plug-in $\widehat{q}_1$ for $c \in (1, c_0]$, c_0 depending of $\widehat{\theta}, \alpha_1, \alpha_2$.

THEOREM. Let $q_{\widehat{\theta}}(\cdot; X) \sim \text{Ga}(\alpha_2, \widehat{\theta}(X))$ be a plug-in density for estimating the density of $Y \sim \text{Ga}(\alpha_2, \theta)$ under KL loss with $\theta \in C = (a, b)$, and based on $X \sim \text{Ga}(\alpha_1, \theta)$. Denote $R(\theta, \widehat{\theta}) = E(\frac{\theta}{\widehat{\theta}(X)} - \log(\frac{\theta}{\widehat{\theta}(X)}) - 1)$ and let $\underline{R} = \inf_{\theta \in C} R(\theta, \widehat{\theta})$. Then, $q_{\widehat{\theta}}(\cdot; X)$ is dominated by $q_{\widehat{\theta},c}(\cdot; X) \sim \text{Ga}(\frac{\alpha_2}{c}, c\widehat{\theta}(X))$ with $1 < c \leq c_0(\underline{R})$, $c_0(\underline{R})$ being the unique solution in $c \in (1, \infty)$ of $G_{\underline{R}}(c) = 0$, with

$$G_s(c) = \alpha_2(\frac{1}{c} - 1)\,(s + 1 - \psi(\alpha_2)) + \frac{\alpha_2}{c}\log c + \log\frac{\Gamma(\frac{\alpha_2}{c})}{\Gamma(\alpha_2)}\,.$$

EXAMPLE. $X \sim p(|x - \theta|), Y \sim q(|y - \theta|)$, L_1 loss, q decreasing and log-concave. KMS (2017) obtain predictive densities

$$\widehat{q}_c(y; x) = \frac{1}{c}\widehat{q}(\frac{|y - x|}{c})$$

that dominate the plug-in $\widehat{q}_1$ for $1 < c \leq c_0$. Important case is normal case.

NO MORE TIME

6. SPHERICALLY SYMMETRIC DISTRIBUTIONS with UNKNOWN LOCATION and SCALE

(a) MODEL

We observe $(X, U) \in \mathbb{R}^{d+k}$, wish to predict $Y \in \mathbb{R}^d$ for the spherically symmetric model density :

$$(X, Y, U) \sim \eta^{d+\frac{k}{2}} f(\eta(\|x-\theta\|^2+\|y-c\theta\|^2+\|u\|^2)$$

with known f, c, unknown $\theta \in \mathbb{R}^d$, $\eta > 0$.

Includes normal case with $X_1, \ldots, X_n, Y$ i.i.d. $N_d(\mu, \sigma^2 I_d)$ Also, scale mixtures of normals with :

$$f(t) = \int_{\mathbb{R}_+} (2\pi z)^{-(d+k/2)} e^{-t/2z} dG(z),$$

with known mixing cdf G.

(b) PREDICTIVE DENSITIES

Based on (x, u), we wish to obtain a predictive density $\hat{q}(\cdot; x, u)$ for the conditional density $Y|x, u$ (simply the marginal density of Y in the normal case) and evaluate its efficiency under Kullback-Leibler loss and risk.

Benchmark predictive density. $\hat{q}_{mre}$, also minimax, Bayes $\hat{q}_{\pi_0}$ with respect to prior density $\pi_0(\theta, \eta) = \frac{1}{\eta}$, is given by a multivariate Student density

$$\hat{q}_{\pi_0}(\cdot; (x, u)) \sim T_d(k, cx, \sqrt{\frac{(1+c^2)\|u\|^2}{k}}) .$$

(Aitchison and Dunsmore, 1975 ; Liang and Barron, 2004 ; for normal case), for all model densities f !

Extension of the class of predictive density estimation improvements for $d \geq 3$ by considering class of alternative densities

$$T_d(k, c\,\hat{\theta}(X, U), \sqrt{\frac{(1+c^2)\|U\|^2}{k}}) .$$

7. CONCLUDING REMARKS

— Informative (defective) properties of plug-in estimators.

— Techniques for improvements include variance expansion and improving the plug-in through a dual loss.

— Different loss functions and models.

SOME REFERENCES

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, **62**, 547-554.

Boisbunon, A. et Maruyama, Y. (2014). Inadmissibility of the best equivariant density in the unknown variance case. *Biometrika*, **101**, 733-740.

Brandwein, A.C., Ralescu, S. and Strawderman, W.E. (1993). Shrinkage estimators of the location parameter for certain spherically symmetric distributions. *AISM*, **45**, 551-565.

Brandwein, A.C and Strawderman, W.E. (1980). Minimax estimation of location parameters for spherically symmetric distributions with concave loss. *AOS*, **8**, 279-284

Brown, L.D., George, E.I., Xu, X. (2008). Admissible predictive density estimation. *Annals of Statistics*, **36**, 1156-1170.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. Studia Sci. Math. Hungar., **2**, 299-318.

DasGupta, A. and Lahiri, S.N. (2012). Density estimation in high and ultra dimensions, regularization, and the $L1$ asymptotics. *Contemporary Developments in Bayesian analysis and Statistical Decision Theory A Festschrift for William E. Strawderman*, IMS Volume Series, **8**, 1-23.

Fourdrinier, D., Marchand, É., Righi, A. & Strawderman, W.E. (2011). On improved predictive density estimation with parametric constraints. *EJS*, **5**, 172-191.

George, E. I., Liang, F. and Xu, X. (2012). From minimax shrinkage estimation to minimax shrinkage prediction. *Stat. Sc.*, **27**, 82-94.

George, E.I., Liang, F., and Xu, X. (2006). Improved minimax predictive densities under Kullback-Leibler loss. *AOS*, **34**, 78-91.

George, E.I. & Xu, X. (2010). Bayesian predictive density estimation. *Frontiers of Statistical Decision Making and Bayesian Analysis. In honor of James O. Berger*, 83-95. Springer.

Kato, K. (2009). Improved prediction for a multivariate normal distribution with unknown mean and variance. *AISM*, **61**, 531-542.

Komaki, F. (2007). Shrinkage priors for Bayesian prediction. *AISM*, **59**, 135-146.

Komaki, F. (2006). Shrinkage priors for Bayesian prediction. *AOS*, **34**, 808-819.

Komaki, F. (2001). A shrinkage predictive distribution for multivariate normal observables. *Biometrika*, **88**, 859-864.

Kubokawa, T., Marchand, É., Strawderman, W.E., Turcotte, J.P. (2013). Minimaxity in predictive density estimation with parametric constraints, *JMA*, **116**, 382-397

Kubokawa, T., Marchand, É. & Strawderman, W.E. (2015). On improved shrinkage estimators under concave loss. Statistics & Probability Letters, 96, 241-246

Kubokawa, T., Marchand, É & Strawderman, W.E. (2017). On predictive density estimation for location families under integrated absolute value loss. *Bernoulli*, **23**, 3197-3212.

Kubokawa, T., Marchand, É. & Strawderman, W.E. (2015). On predictive density estimation for location families under integrated L_2 loss. *JMA*, **142**, 57-74.

Lawless, J.F. and Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika*, **92**, 529-542.

Liang, F. and Barron, A. (2004). Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE T. Inf. Theo.*, **50**, 2708-2726.

L'Moudden, A., Marchand, É., Kortbi, O. & Strawderman, W.E. (2017). On predictive density estimation for Gamma models with parametric constraints. *Journal of Statistical Planning and Inference*, **185**, 56-68.

Maruyama, Y. (2003). A robust generalized Bayes estimator improving on the James-Stein estimator for spherically symmetric distributions. *Statistics & Decisions*, 1-9.

Murray, G.D. (1977). A note on the estimation of probability density functions. *Biome*
64, 150-152.