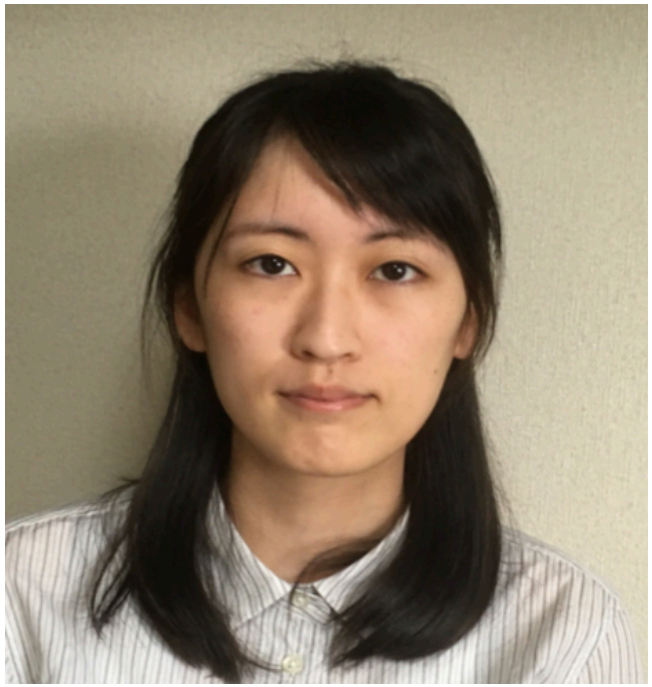
A Class of Tractable Predictive Densities for Curved Exponential Families

Fumiyasu Komaki

The University of Tokyo, RIKEN CBS

Symposium in Memory of Charles Stein
June 18, 2019

This talk is based on a joint work with Michiko Okudo.



A statistical model of a curved exponential family

$$\mathscr{P} = \Big\{ p(x;u) = s(x)\exp(\theta^i(u)x_i - \Psi(\theta(u))) \;\Big|\; u := (u^a) \in U,\ a = 1,\dots,d,\ i = 1,\dots,m \Big\},$$

where $U \subset \mathbb{R}^d$ and $1 \le d \le m$.

We have observations $x^n = \{x(1), x(2), \dots, x(n)\}$ independently distributed according to a density $p(x;u) \in \mathscr{P}$.

Einstein's summation convention: if an index occurs as an upper and lower index in one term, then the summation is implied.

$$\theta^i x_i = \sum_{i=1}^{m} \theta^i x_i.$$

The objective is to construct a predictive density of $y = x(n+1)$ which is independently distributed according to the same density $p(y;u)$.

The Kullback–Leibler divergence loss

$$D\{p(y;u);\hat{p}(y;u)\} = \int p(y;u) \log \frac{p(y;u)}{\hat{p}(y;u)} dy.$$

The risk function

$$\mathrm{E}[D\{p(y;u);\hat{p}(y;u)\}] = \int p(x^n;u) D\{p(y;u);\hat{p}(y;u)\} dx^n.$$

The Bayes risk

$$\int \pi(u) \int p(x^n;u) D\{p(y;u);\hat{p}(y;u)\} dx^n du$$

with respect to a prior $\pi(u)$.

Two methods for constructing predictive densities

(i) A plugin density

$$p(y; \hat{u}(x^n))$$

is constructed by plugging-in an estimator $\hat{u}(x^n)$ to the unknown parameter of the model, which is included in $\mathscr{P}$.

(ii) A Bayesian predictive density is defined by

$$p_\pi(y \mid x^n) = \int p(y; u) p_\pi(u \mid x^n) du$$

where $p_\pi(u \mid x^n)$ is the posterior density

$$p_\pi(u \mid x^n) = \frac{p(x^n; u)\pi(u)}{\int p(x^n; u)\pi(u) du}$$

of u.

- In many examples, the Bayesian predictive density is not included in the model $\mathscr{P}$.
- The Bayesian predictive density is optimal about the Bayes risk in terms of the Kullback–Leibler divergence in the family of all probability distributions, which we denote as $\mathscr{F}$ (Aitchison, 1975).
- However, the explicit form of the Bayesian predictive density is often intractable.
- In such examples, the numerical calculations of Bayesian predictive densities are burdensome because it involves integrations in the space of probability density functions.

For multivariate normal models, Xu and Zhou (2011) proposed a class of empirical Bayes predictive densities to avoid intractable implementation of Bayesian predictive densities.

Here, we consider a distinct class of predictive densities and our focus is on models of curved exponential families including full exponential families such as multivariate normal models.

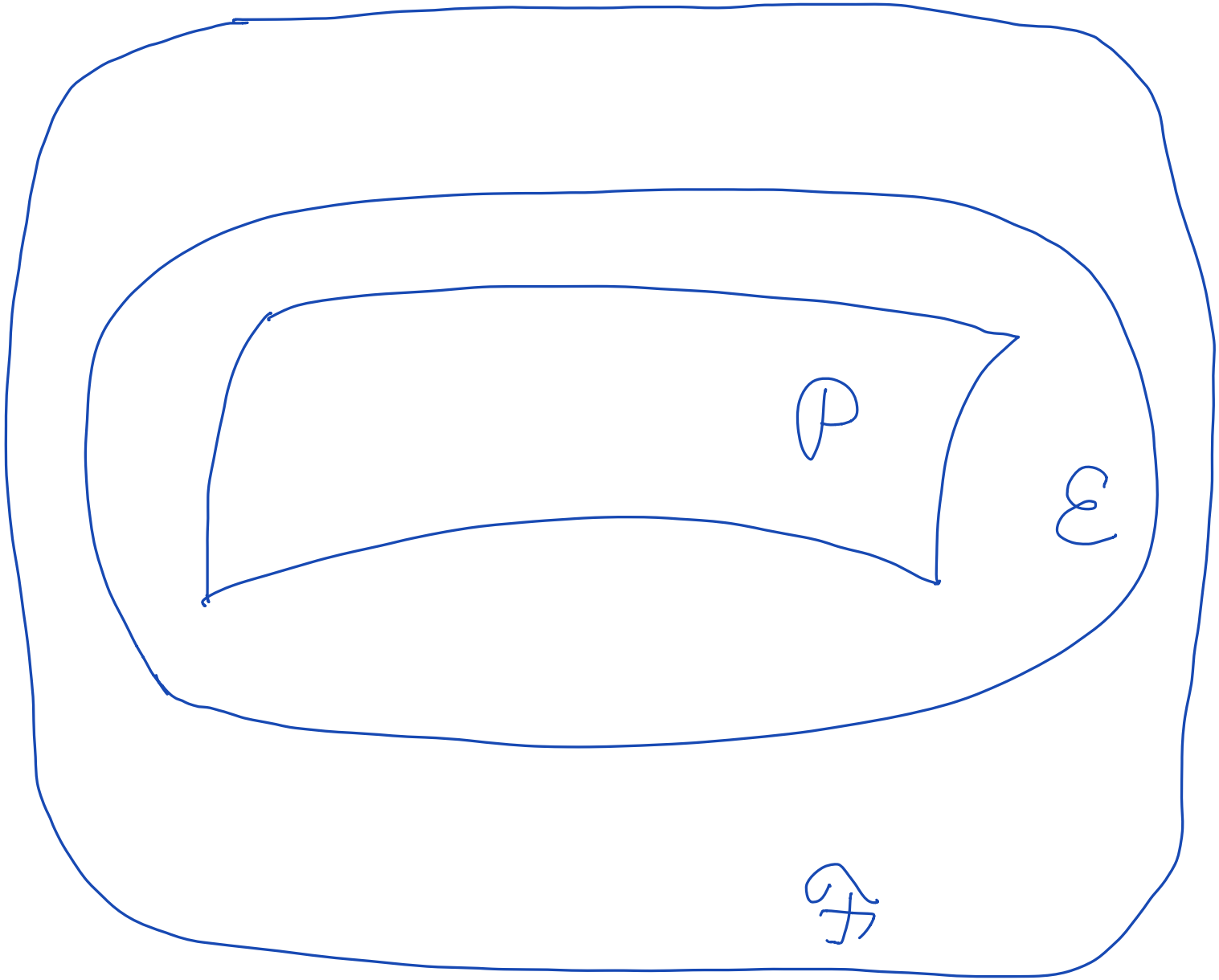
We consider a full exponential family

$$\mathscr{E} = \Bigg\{ p(x;\theta) = s(x)\exp(\theta^i x_i - \Psi(\theta)) \\ \Bigg| \ \theta := (\theta^i) \in \Theta \subseteq \mathbb{R}^m,\ i = 1,\ldots,m \Bigg\}$$

including $\mathscr{P}$.

We refer to plugin distributions in $\mathscr{E}$ as extended plugin distributions, and investigate the properties of extended plugin distributions.

The inclusion relation is $\mathscr{P} \subseteq \mathscr{E} \subseteq \mathscr{F}$ and we consider the middle layer of the three-layer structure.


P
E
U

The coordinate system $\theta = (\theta^i)$ $(i = 1, \ldots, m)$ is called natural parameters of exponential families.

Another coordinate system $\eta = (\eta_i)$ defined by

$$\eta_i = E[x_i] = \frac{\partial}{\partial \theta^i} \Psi(\theta) \qquad (i = 1, \ldots, m),$$

is called expectation parameters.

Outline (1/2)

- We show that the extended plugin density with the posterior mean of η is optimal in the class of the extended plugin densities.
- The Bayes risk of the Bayesian predictive density is not greater than that of the extended plugin density because the Bayesian predictive density is optimal under the Bayes risk with the Kullback–Leibler loss,
- However, evaluating extended plugin densities is less difficult than evaluating Bayesian predictive densities.
- The properties of extended plugin densities are investigated from an information-geometric view.

Outline (2/2)

- The extended plugin of the Bayes estimator and the Bayesian predictive distribution are parallel in that the Bayesian predictive distribution is optimal in $\mathscr{F}$ and the extended plugin of the Bayes estimator denoted herein as $p(y;\hat{\eta}_\pi)$ is optimal in $\mathscr{E}$ regarding the Bayes risk.
- We show that several geometric results of Bayesian predictive densities also hold for $p(y;\hat{\eta}_\pi)$ parallelly.

$\mathcal{F}$: all probability densities

$p(x)$: m-representation

$\log p(x)$: e-representation

infinite dimensional model

$\mathcal{E}$: full exponential family

η : m-affine parameter

θ : e-affine parameter

finite dimensional model

1. Introduction
2. Extended plugin densities
3. Information geometry of extended plugin densities
4. Shrinkage priors

To construct predictive densities, plugin densities with estimators of u (e.g., the maximum likelihood estimator $\hat{u}_{\mathrm{MLE}}$ or the Bayes estimator $\hat{u}_\pi$) are often used.

The posterior mean of u for plugin densities is

$$\bar{u}_\pi := \frac{\int u p(x^n; u)\pi(du)}{\int p(x^n; u)\pi(du)},$$

and when we consider the full exponential family, the posterior mean of η for extended plugin densities is

$$\bar{\eta}_\pi := \frac{\int \eta p(x^n; \eta(u))\pi(du)}{\int p(x^n; \eta(u))\pi(du)}.$$

Note that $\bar{\eta}_\pi(x^n) \neq \eta(\bar{u}_\pi(x^n))$ in general.

Example (spike model)

We consider inference of the eigenvector and the eigenvalue of the I-dimensional Gaussian distribution $\mathrm{N}(0, \Sigma)$ with zero mean vector and unknown covariance matrix.

The covariance matrix Σ is supposed to be

$$\Sigma = \lambda \omega \omega^{\top} + I_I,$$

where the vector $\omega \in \mathbb{R}^I$ satisfies $\omega^{\top}\omega = 1$ and $\lambda > 0$.

The eigenvalues of the matrix Σ are $\lambda + 1, 1, \ldots, 1$, and ω is the first eigenvector.

The model $\mathscr{P} = \{\mathrm{N}(0, \Sigma) \mid (\omega, \lambda)\}$ is parametrized by (ω, λ), and the plugin density with the posterior mean $\bar{\omega}_\pi, \bar{\lambda}_\pi$ is $\mathrm{N}(0, \Sigma(\bar{\omega}_\pi, \bar{\lambda}_\pi))$.

On the other hand, the posterior mean of the matrix Σ is another natural estimator of Σ other than $\Sigma(\bar{\omega}_\pi, \bar{\lambda}_\pi)$.

In principal component analysis, estimating the matrix and then decomposing the estimated matrix is a natural way to obtain estimators of eigenvalues and eigenvectors.

The components of Σ comprise the coordinate system (η_i) of the extended statistical model $\mathscr{E} = \{\mathrm{N}(0, \Sigma) \mid \Sigma\}$, thus $p(y; \bar{\Sigma}_\pi)$ is the extended plugin density with the posterior mean of η.

Consequently, $\bar{\eta}_\pi$ appears to be a natural estimator for curved exponential families.

The optimal extended plugin density

Proposition

The Bayes risk of $p(y;\hat{\eta})$, where $\hat{\eta}$ is an estimator of η, is minimized when $\hat{\eta} = \bar{\eta}_\pi$.

Proof

Let $\hat{\theta}$ be an estimator of θ. Note that θ and η are functions of u. The Kullback–Leibler loss of $p(y;\hat{\theta})$ is

$$D\{p(y;\theta(u)); p(y;\hat{\theta})\} = \int p(y;\theta) \log\left(\frac{\exp(\theta^i y_i - \Psi(\theta))}{\exp(\hat{\theta}^i y_i - \Psi(\hat{\theta}))}\right) dy$$
$$= (\theta^i - \hat{\theta}^i)\eta_i - (\Psi(\theta) - \Psi(\hat{\theta})).$$

Hence

$$\int p(u \mid x^n) D\{p(y;\theta(\eta)); p(y;\hat{\theta}(\eta))\} du$$
$$= D\{p(y;\overline{\eta}); p(y;\hat{\theta})\} + \left(-\theta^i(\overline{\eta})\overline{\eta}^i + \overline{\theta^i \eta_i} + \Psi(\theta(\overline{\eta})) - \overline{\Psi(\theta)}\right),$$

where, for a function $f(\eta)$, $\overline{f(\eta)} = \int p(u \mid x^n) f(\eta) du$.

The Bayes risk is minimized when $\hat{\theta} = \theta(\overline{\eta}) = \theta(\overline{\eta}_\pi)$.

Thus $p(y \mid \bar{\eta}_\pi)$ is optimal with respect to the Bayes risk in the class of extended plugin densities. □

Example Fisher circle model (1/4)

Two dimensional Gaussian distributions $\mathrm{N}(\mu, I_2)$ with unknown mean vector μ and the identity covariance matrix I_2.

The probability density is

$$\begin{aligned} p(x;\mu) &= \frac{1}{2\pi}\exp\left[-\frac{1}{2}\left\{(x_1-\mu_1)^2+(x_2-\mu_2)^2\right\}\right] \\ &= \frac{1}{2\pi}\exp\left(-\frac{1}{2}(x_1^2+x_2^2)\right)\exp\left(x_1\mu_1+x_2\mu_2-\frac{1}{2}\left(\mu_1^2+\mu_2^2\right)\right). \end{aligned}$$

When the mean vector μ is expressed as

$$\mu_1 = \cos\omega, \mu_2 = \sin\omega,$$

the 1-dimensional submodel is called the Fisher circle model.

Here,

$$\theta^1 = \eta_1 = \mu_1,\ \theta^2 = \eta_2 = \mu_2.$$

Example Fisher circle model (2/4)

We derive the Bayes estimator of ω, the extended plugin of the Bayes estimator $\hat{\eta}_\pi$, and the Bayesian predictive density.

For $x^n = \{x(1), x(2), \ldots, x(n)\}$,

$$
\begin{aligned}
p(x^n; \omega) &= \prod_{t=1}^{n} \frac{1}{2\pi} \exp\left(-\frac{\|x(t) - \mu(\omega)\|^2}{2} \right) \\
&= \frac{1}{(2\pi)^n} \exp\left(-\frac{\sum_{t=1}^{n} (x_1(t)^2 + x_2(t)^2)}{2} + \frac{n}{2}(\bar{x}_1^2 + \bar{x}_2^2) \right) \\
&\quad \times \exp\left(-\frac{n}{2} \|\bar{x} - \mu(\omega)\|^2 \right)
\end{aligned}
$$

where $\bar{x} = \sum_{t=1}^{n} x(t)/n$.

Example Fisher circle model (3/4)

Let $\bar{x} = (\|\bar{x}\| \cos\phi, \|\bar{x}\| \sin\phi)^\top$.

When the uniform prior

$$\pi(\omega) \propto 1$$

is adopted, the posterior distribution is

$$p_\pi(\omega \mid x^n) = \frac{1}{2\pi I_0(n\|\bar{x}\|)} \exp(n\|\bar{x}\| \cos(\omega - \phi)).$$

The Bayes estimator

$$\hat{\omega}_\pi = \phi.$$

Thus,

$$\eta(\hat{\omega}_\pi) = \bar{x}/\|\bar{x}\|.$$

Example Fisher circle model (4/4)

The posterior mean of η is

$$\hat{\eta}_\pi = \frac{I_1(n\|\bar{x}\|)}{I_0(n\|\bar{x}\|)} \frac{\bar{x}}{\|\bar{x}\|},$$

which is not included in the circle parametrized by ω and $\hat{\eta}_\pi \neq \eta(\hat{\omega}_\pi)$.

The Bayesian predictive density is given by

$$p_\pi(y \mid x^n) = \frac{1}{2\pi} \frac{I_0(\|y + n\bar{x}\|)}{I_0(n\|\bar{x}\|)} \exp\left\{-\frac{1}{2}(\|y\|^2 + 1)\right\}.$$

Therefore $p_\pi(y \mid x^n)$ is not included in $\mathscr{P}$ or $\mathscr{E}$ because it is not a two dimensional Gaussian with covariance matrix I_2.

Information-geometric notions

Let $a, b, \ldots$ be indices for u.

Let $T_u\mathscr{P}$ be the tangent space of $\mathscr{P}$ at a point u.

The tangent space $T_u\mathscr{P}$ is identified with the vector space spanned by $\partial_a p(x;u)$ $(a = 1, \ldots, d)$, where ∂_a denotes $\partial/\partial u^a$.

Define inner products in the tangent space by

$$\langle r, s \rangle = \int rs \frac{1}{p(x;u)} dx. \tag{1}$$

For a statistical model $\mathscr{P}$, each component of the Fisher information matrix is defined by

$$g_{ab}(u) = \langle \partial_a p(x;u), \partial_b p(x;u) \rangle = \int \frac{\partial_a p(x;u) \partial_b p(x;u)}{p(x;u)} dx$$

and let g^{ab} be a component of the inverse matrix of (g_{ab}).

The e-connection coefficients and the m-connection coefficients are defined by

$$\overset{\mathrm{e}}{\Gamma}_{abc}(u) = \int p(x;u)\{\partial_a\partial_b \log p(x;u)\}\{\partial_c \log p(x;u)\}dx$$

and

$$\overset{\mathrm{m}}{\Gamma}_{abc}(u) = \int \frac{\partial_a\partial_b p(x;u)\partial_c p(x;u)}{p(x;u)}dx,$$

respectively. We define

$$\overset{\mathrm{e}}{\Gamma}_{ab}^{\ c} := \overset{\mathrm{e}}{\Gamma}_{abd}g^{dc}, \quad \overset{\mathrm{m}}{\Gamma}_{ab}^{\ c} := \overset{\mathrm{m}}{\Gamma}_{abd}g^{dc},$$

and

$$\begin{aligned} T_{abc} &= \overset{\mathrm{m}}{\Gamma}_{abc} - \overset{\mathrm{e}}{\Gamma}_{abc} \\ &= \int p(x;u)\{\partial_a \log p(x;u)\}\{\partial_b \log p(x;u)\}\{\partial_c \log p(x;u)\}dx. \end{aligned}$$

The Jeffreys prior density is given by

$$\pi_J(u) = \sqrt{|g(u)|},$$

where $|g(u)|$ is the determinant of the matrix $(g_{ab}(u))$.

The coordinate systems (θ^i) and (η_i) of the exponential family $\mathscr{E}$ are dual to each other in the sense that

$$\left\langle \frac{\partial}{\partial \theta^i} p(x;\theta), \frac{\partial}{\partial \eta_j} p(x;\theta) \right\rangle = \delta^i_j \tag{2}$$

where δ^i_j is the Kronecker delta.

For a curved exponential family, e-connection coefficients and m-connection coefficients are expressed as

$$\overset{\mathrm{e}}{\Gamma}_{abc} = (\partial_a \partial_b \theta^i)(\partial_c \eta_i) \text{ and } \overset{\mathrm{m}}{\Gamma}_{abc} = (\partial_a \partial_b \eta_i)(\partial_c \theta^i), \tag{3}$$

respectively.

The asymptotic expansion of $\hat{\eta}_\pi$ around $\eta(\hat{u}_{\mathrm{MLE}})$.

Theorem

The posterior mean of η, which is the Bayes estimator of η, based on a prior $\pi(u)$ is expanded as

$$\hat{\eta}_\pi = \eta(\hat{u}_{\mathrm{MLE}}) + \frac{g^{ab}(\hat{u}_{\mathrm{MLE}})}{2n}\left(\partial_a\partial_b\eta(\hat{u}_{\mathrm{MLE}}) - \overset{\mathrm{m}}{\Gamma}{}_{ab}^{c}(\hat{u}_{\mathrm{MLE}})\partial_c\eta(\hat{u}_{\mathrm{MLE}})\right)$$
$$+ \frac{g^{ab}(\hat{u}_{\mathrm{MLE}})}{n}\left(\partial_b \log\frac{\pi}{\pi_J}(\hat{u}_{\mathrm{MLE}}) + \frac{T_b(\hat{u}_{\mathrm{MLE}})}{2}\right)\partial_a\eta(\hat{u}_{\mathrm{MLE}}) + \mathrm{o}(n^{-1}),$$

where $T_a = T_{abc}g^{bc}$.

Let $L(\eta) = \frac{1}{n}\sum_{t=1}^{n} \log p(x(t);\eta)$. Then, $\hat{\eta}_\pi(x^n) = \{(\hat{\eta}_\pi)_i(x^n)\}$ is given by

$$\hat{\eta}_\pi(x^n) = \frac{\int \eta(u) p(x^n;\eta(u))\pi(u)du}{\int p(x^n;\eta(u))\pi(u)du} = \frac{\int \eta(u)\exp(nL(\eta(u)))\pi(u)du}{\int \exp(nL(\eta(u)))\pi(u)du}.$$

We approximate $\hat{\eta}_\pi$ by the Laplace method.

The asymptotic expansion of the extended plugin density

Theorem

The extended plugin density with $\hat{\eta}_\pi$ is expanded as

$$
\begin{aligned}
p(y;\hat{\eta}_\pi) &= p(y;\eta(\hat{u}_{\mathrm{MLE}})) \\
&+ \frac{g^{ab}(\hat{u}_{\mathrm{MLE}})}{2n}\left(\partial_a\partial_b\eta_i(\hat{u}_{\mathrm{MLE}}) - \overset{\mathrm{m}}{\Gamma}{}^{c}_{ab}(\hat{u}_{\mathrm{MLE}})\partial_c\eta_i(\hat{u}_{\mathrm{MLE}})\right)\partial^i p(y;\hat{u}_{\mathrm{MLE}}) \\
&+ \frac{g^{ab}(\hat{u}_{\mathrm{MLE}})}{n}\left(\partial_b \log\frac{\pi}{\pi_J}(\hat{u}_{\mathrm{MLE}}) + \frac{T_b(\hat{u}_{\mathrm{MLE}})}{2}\right)\partial_a p(y;\hat{u}_{\mathrm{MLE}}) + \mathrm{o}_p(n^{-1}),
\end{aligned}
$$

where $\hat{u}_{\mathrm{MLE}}$ is the maximum likelihood estimator and $\partial^i = \frac{\partial}{\partial\eta_i}$.

K (1996) gave an asymptotic expansion of Bayesian predictive distributions around the plugin of $\hat{u}_{\mathrm{MLE}}$, and our results are parallel to the results.

The shift to $p(y;\hat{\eta}_\pi)$ in Theorem 2 is composed of two components, one "parallel" to and the other "orthogonal" to the model $\mathscr{P}$.

That is, the term

$$\frac{g^{ab}(\hat{u}_{\mathrm{MLE}})}{n}\left(\partial_b \log \frac{\pi}{\pi_J}(\hat{u}_{\mathrm{MLE}}) + \frac{T_b(\hat{u}_{\mathrm{MLE}})}{2}\right)\partial_a p(y;\hat{u}_{\mathrm{MLE}})$$

is included in the space spanned by $\partial_a p(y;\eta)$ $(a = 1,\ldots,d)$ and the term

$$\frac{g^{ab}(\hat{u}_{\mathrm{MLE}})}{2n}\left(\partial_a\partial_b\eta_i(\hat{u}_{\mathrm{MLE}}) - \overset{\mathrm{m}}{\Gamma}{}_{ab}^{c}(\hat{u}_{\mathrm{MLE}})\partial_c\eta_i(\hat{u}_{\mathrm{MLE}})\right)\partial^i p(y;\hat{u}_{\mathrm{MLE}})$$

is orthogonal to $\partial_a p(x;\eta)$ $(a = 1,\ldots,d)$ with respect to the inner product (1).

$$
\begin{aligned}
&\left\langle \left(\partial_a \partial_b \eta_i - \overset{\mathrm{m}}{\Gamma}{}_{ab}^{c} \partial_c \eta_i\right) \partial^i p(y;\eta), \partial_e p(y;\eta) \right\rangle \\
&\quad = \int \partial_a \partial_b \eta_i \frac{\partial p(y;\eta)}{\partial \eta_i} \frac{\partial \theta^j}{\partial u^e} \frac{\partial p(y;\eta)}{\partial \theta^j} \frac{1}{p(y;\eta)} dy - \overset{\mathrm{m}}{\Gamma}{}_{ab}^{c} g_{ce} \\
&\quad = \partial_a \partial_b \eta_i \partial_e \theta^i - \overset{\mathrm{m}}{\Gamma}{}_{abe} = 0.
\end{aligned}
$$

We devide the tangent vectors of $\mathscr{E}$ at η into two parts, namely those parallel to $\mathscr{P}$ and those orthogonal to $\mathscr{P}$.

For each point $\eta \in \mathscr{E}$, the tangent space $T_\eta \mathscr{E}$ is identified with the vector space spanned by

$$\frac{\partial}{\partial \eta_i} p(x;\eta) \ (i = 1, \ldots, m).$$

The tangent space $T_u \mathscr{P}$ is a subspace of $T_\eta \mathscr{E}$.

Let $A(u)$ be an $(m-d)$-dimensional smooth submanifold of $\mathscr{E}$ attached to each point $u \in \mathscr{P}$ and assume that $A(u)$ orthogonally transverses $\mathscr{E}$ at $\eta(u)$.

Such a family of submanifolds $\{A(u)\}$ is called an ancillary family.

We introduce an adequate coordinate system $v = (v^\kappa)$ $(\kappa = d+1, \ldots, m)$ to $A(u)$ so that a pair (u, v) uniquely specifies a point of $\mathscr{E}$ in the neighborhood of $\eta(u)$.

We adopt a coordinate system v on $A(u)$ such that $\eta(u, v) \in \mathscr{M}$ if $v = 0$.

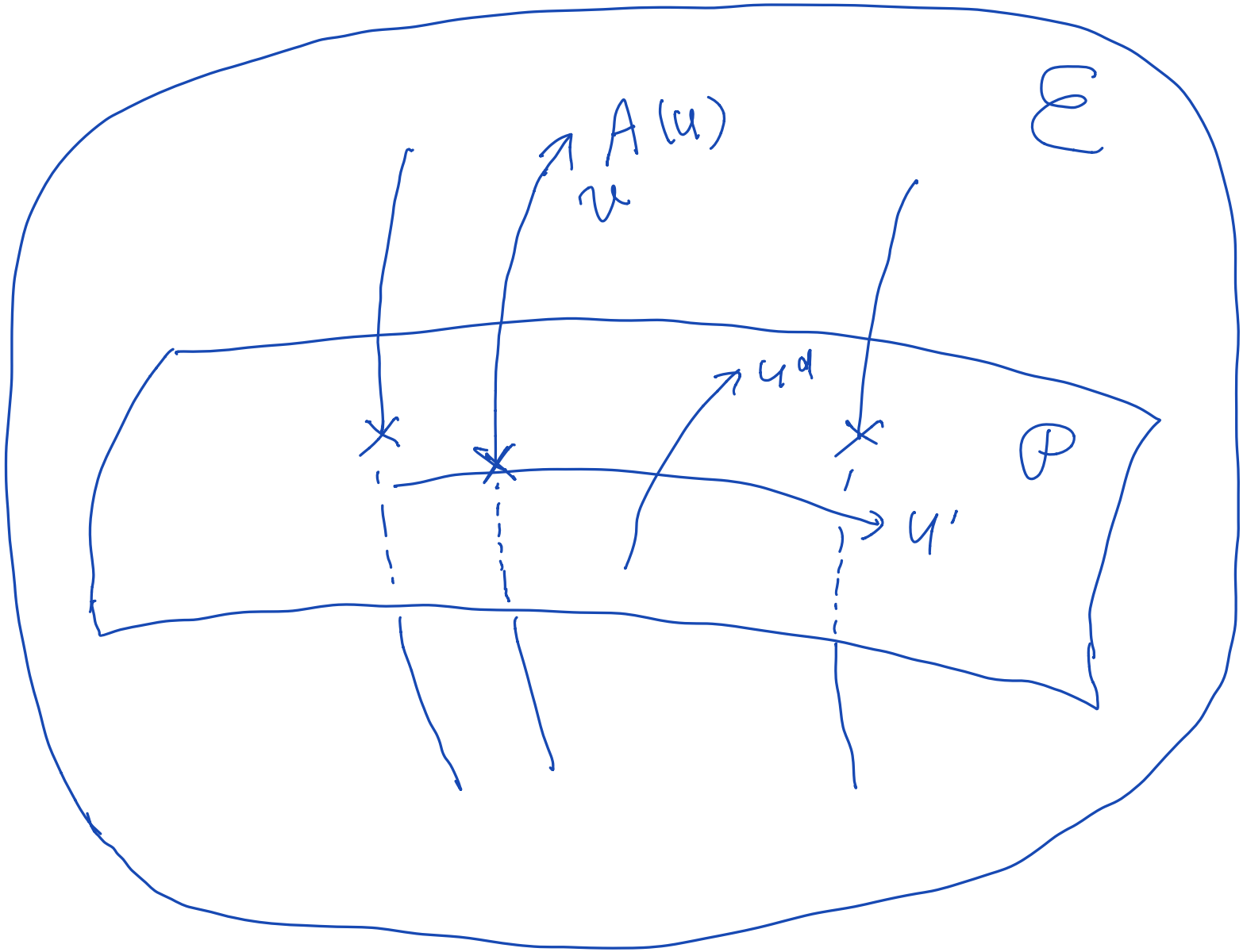
Then, we have

$$\operatorname{span}\left\{\partial^i p(x;\eta)\right\} = \operatorname{span}\left\{\partial_a p(x;\eta), \partial_\kappa p(x;\eta)\right\}$$

where $\partial_\kappa p(x;\eta) = \frac{\partial}{\partial v^\kappa} p(x;\eta)$.

Since $A(u)$ orthogonally transverses $\mathscr{E}$, we have

$$\langle \partial_a p(x;\eta), \partial_\kappa p(x;\eta) \rangle = 0 \quad (a = 1, \ldots, d,\ \kappa = d+1, \ldots, m).$$


E
A(u)
u
u'
P

In the following discussion, we consider extended plugin densities $p(y;\hat{\eta})$ with estimators $\hat{\eta} = \eta(\hat{u},\hat{v})$ where $\hat{u},\hat{v}$ can be expressed in the form

$$\hat{u} = \hat{u}_{\mathrm{MLE}} + \frac{1}{n}\alpha(\hat{u}_{\mathrm{MLE}}) + \mathrm{o}_p(n^{-1}), \quad \hat{v} = \frac{1}{n}\beta(\hat{u}_{\mathrm{MLE}}) + \mathrm{o}_p(n^{-1}),$$

respectively.

Here $\alpha^a(u), \beta^\kappa(u)$ are smooth functions of $\mathrm{O}_p(1)$.

Those densities can be expanded as

$$p_{\alpha,\beta}(y;\hat{u}_{\mathrm{MLE}}) = p(y;\hat{u}_{\mathrm{MLE}}) + \frac{1}{n}\alpha^a(\hat{u}_{\mathrm{MLE}})\partial_a p(y;\eta(\hat{u}_{\mathrm{MLE}})) + \frac{1}{n}\beta^\kappa(\hat{u}_{\mathrm{MLE}})\partial_\kappa p(y;\eta(\hat{u}_{\mathrm{MLE}})) + \mathrm{o}_p(n^{-1}).$$

This class of extended plugin densities include $p(x; \hat{u}_{\mathrm{MLE}})$ and $p(x; \hat{\eta}_\pi)$.

For $\hat{u} = \hat{u}_{\mathrm{MLE}}$ and $\beta^\kappa(u) = 0$ for $\kappa = d+1, \ldots, m$, the density is the plugin density with the maximum likelihood estimator $\hat{u}_{\mathrm{MLE}}$.

The extended plugin density with the Bayes estimator $\hat{\eta}_\pi$ in Theorem 1 is given by

$$\alpha^a(\hat{u}_{\mathrm{MLE}}) = g^{ab}(\hat{u}_{\mathrm{MLE}})\left(\partial_b \log \frac{\pi}{\pi_J} + \frac{T_b}{2}\right), \quad \beta^\kappa = \frac{1}{2} \overset{\mathrm{m}}{H}{}_{ab}^{\kappa}(\hat{u}_{\mathrm{MLE}}) g^{ab}(\hat{u}_{\mathrm{MLE}})$$

where

$$\overset{\mathrm{m}}{H}_{ab\kappa} = \langle \partial_a \partial_b p(x; u), \partial_\kappa p(x; \eta) \rangle = (\partial_a \partial_b \eta_i)(\partial_\kappa \theta^i)$$

is the mixture embedding curvature of $\mathscr{P}$ in $\mathscr{E}$ and $\overset{\mathrm{m}}{H}{}_{ab}^{\kappa} = \overset{\mathrm{m}}{H}_{ab\lambda} g^{\kappa\lambda}$.

This can be confirmed as follows. Let

$$h_{ab} = (\partial_a \partial_b \eta_i(u))\partial^i p(x;\eta) - \overset{\mathrm{m}}{\Gamma}{}_{ab}^{c} \partial_c p(x;u),$$

then the orthogonal component in Theorem 1 is

$$\frac{g^{ab}(\hat{u}_{\mathrm{MLE}})}{2n} h_{ab}(\hat{u}_{\mathrm{MLE}}).$$

Since h_{ab} $(a, b = 1, \ldots, d)$ are included in the space spanned by $\partial_\kappa p(x;u)$ $(\kappa = d+1, \ldots, m)$,

$$\begin{aligned} h_{ab} &= \langle h_{ab}, \partial_\lambda p(x;\eta) \rangle g^{\kappa\lambda} \partial_\kappa p(x;\eta) \\ &= \left\langle \partial_a \partial_b \eta_i(u) \partial^i p(x;\eta) - \overset{\mathrm{m}}{\Gamma}{}_{ab}^{c} \partial_c p(x;u), \partial_\lambda p(x;\eta) \right\rangle g^{\kappa\lambda} \partial_\kappa p(x;\eta). \end{aligned}$$

Because $\langle \partial_c p(x;\eta), \partial_\lambda p(x;\eta) \rangle = 0$,

$$
\begin{aligned}
h_{ab} &= \langle \partial_a \partial_b \eta_i(u) \partial^i p(x;\eta), \partial_\lambda p(x;u) \rangle g^{\kappa\lambda} \partial_\kappa p(x;\eta) \\
&= \partial_a \partial_b \eta_i(u) \left\langle \partial^i p(x;u), \frac{\partial \theta^j}{\partial v^\lambda} \frac{\partial p(x;\eta)}{\partial \theta^j} \right\rangle g^{\kappa\lambda} \partial_\kappa p(x;\eta) \\
&= \partial_a \partial_b \eta_i(u) \partial_\lambda \theta^i g^{\kappa\lambda} \partial_\kappa p(x;\eta) \\
&= \overset{\mathrm{m}}{H}{}^{\kappa}_{ab} \partial_\kappa p(x;\eta).
\end{aligned}
$$

Optimal orthogonal shift

We derive the Kullback–Leibler risk of the extended plugin densities.

Proposition

The Kullback–Leibler risk of a extended plugin density $p_{\alpha,\beta}(y; \hat{u}_{\mathrm{MLE}})$ *is asymptotically expanded as*

$$
\begin{aligned}
&E[D\{p(y;u), p_{\alpha,\beta}(y;\hat{u}_{\mathrm{MLE}})\}] \\
&= \frac{1}{2n^2} g_{ab}(u)\alpha^a(u)\alpha^b(u) + \frac{1}{n^2}\overset{\mathrm{e}}{\nabla}_a \alpha^a(u) \\
&\quad + \frac{1}{2n^2} g_{\kappa\lambda}(u)\beta^\kappa(u)\beta^\lambda(u) - \frac{1}{2n^2}\overset{\mathrm{m}}{H}_{ab\kappa}(u) g^{ab}(u)\beta^\kappa(u) \\
&\quad + (\text{terms independent of } \alpha, \beta) + \mathrm{o}(n^{-2}),
\end{aligned}
$$

where $\overset{\mathrm{e}}{\nabla}_a \alpha^b = \partial_a \alpha^b + \overset{\mathrm{e}}{\Gamma}_{ac}^{\ b} \alpha^c$.

We obtain the optimal orthogonal shift.

Theorem

The optimal β^κ is given by

$$\beta^\kappa_{\mathrm{opt}}(u) = \frac{1}{2} \overset{\mathrm{m}}{H}{}^\kappa_{ab}(u) g^{ab}(u).$$

Proof.

The risk in Proposition 2 is

$$
\begin{aligned}
&\mathrm{E}[D\{p(y;u), p_{\alpha,\beta}(y;\hat{u}_{\mathrm{MLE}})\}] \\
=&\frac{1}{2n^2} g_{ab}\alpha^a\alpha^b + \frac{1}{n^2}\overset{\mathrm{e}}{\nabla}_a \alpha^a(u) \\
&+ \frac{1}{2n^2} g_{\kappa\lambda}\left(\beta^\lambda - \frac{1}{2}\overset{\mathrm{m}}{H}{}_{ab}^{\ \lambda} g^{ab}\right)\left(\beta^\kappa - \frac{1}{2}\overset{\mathrm{m}}{H}{}_{cd}^{\ \kappa} g^{cd}\right) - \frac{1}{8n^2}\overset{\mathrm{m}}{H}{}_{ab}^{\ \lambda}\overset{\mathrm{m}}{H}{}_{cd}^{\ \kappa} g^{ab} g^{cd} g_{\kappa\lambda} \\
&+ (\text{terms independent of } \alpha, \beta) + \mathrm{o}(n^{-2}).
\end{aligned}
$$

Thus β is optimal when

$$
\beta^\kappa(u) = \frac{1}{2}\overset{\mathrm{m}}{H}{}_{ab}^{\ \kappa}(u) g^{ab}(u).
$$

□

Therefore, the orthogonal component of the shift in Theorem 2 is optimal.

The extended plugin density with $\hat{\eta}_\pi$ has the optimal shift.

The risk difference between a plugin density $p_{\alpha,0}(y; \hat{u}_{\mathrm{MLE}}))$ and a extended plugin density $p_{\alpha,\beta_{\mathrm{opt}}}(y; \hat{u}_{\mathrm{MLE}}))$ with the optimal orthogonal shift is given by

$$\begin{aligned}&\mathrm{E}[D(p(y;u), p_{\alpha,0}(y;\hat{u}_{\mathrm{MLE}}))] - \mathrm{E}[D(p(y;u), p_{\alpha,\beta_{\mathrm{opt}}}(y;\hat{u}_{\mathrm{MLE}}))] \\ =&\frac{1}{8n^2} \overset{\mathrm{m}}{H}{}_{ab}^{\lambda} \overset{\mathrm{m}}{H}{}_{cd}^{\kappa} g^{ab} g^{cd} g_{\kappa\lambda} + \mathrm{o}(n^{-2}).\end{aligned}$$

Here, $\overset{\mathrm{m}}{H}{}_{ab}^{\lambda} \overset{\mathrm{m}}{H}{}_{cd}^{\kappa} g^{ab} g^{cd} g_{\kappa\lambda}$ is the mixture mean curvature of $\mathscr{P}$ embedded in $\mathscr{M}$ at u_0.

The optimal orthogonal obtained by K (1996) is not in the tangent space of exponential families $\mathscr{E}$.

Thus, the shifted plugin density is not icncluded in $\mathscr{E}$.

Our optimal shift is included in the tangent space of $\mathscr{E}$, and the shifted plugin density is included in $\mathscr{E}$.

The optimal orthogonal shift of K (1996) is

$$\frac{g^{ab}(\hat{u}_{\mathrm{MLE}})}{2n}\partial_a\partial_b p(y;\hat{u}_{\mathrm{MLE}}) - \overset{\mathrm{m}}{\Gamma}{}_{ab}^{c}(\hat{u}_{\mathrm{MLE}})\partial_c p(y;\hat{u}_{\mathrm{MLE}}). \tag{4}$$

Our optimal shift is a projection of (4) onto the tangent space of $\mathscr{E}$ because

$$\left\langle \frac{g^{ab}(\hat{u}_{\mathrm{MLE}})}{2n}\partial_a\partial_b p(y;\hat{u}_{\mathrm{MLE}}) - \overset{\mathrm{m}}{\Gamma}{}_{ab}^{c}(\hat{u}_{\mathrm{MLE}})\partial_c p(y;\hat{u}_{\mathrm{MLE}}), \partial_\lambda p(y;\hat{u}_{\mathrm{MLE}}) \right\rangle g^{\lambda\kappa}\partial_\kappa p(y;\hat{u}_{\mathrm{MLE}})$$

$$= \frac{g^{ab}(\hat{u}_{\mathrm{MLE}})}{2n}\left\langle \partial_a\partial_b p(y;\hat{u}_{\mathrm{MLE}}), \partial_\lambda p(y;\hat{u}_{\mathrm{MLE}}) \right\rangle g^{\kappa\lambda}\partial_\kappa p(y;\hat{u}_{\mathrm{MLE}})$$

$$= \frac{g^{ab}(\hat{u}_{\mathrm{MLE}})}{2n}\overset{\mathrm{m}}{H}{}_{ab}^{\kappa}\partial_\kappa p(y;\hat{u}_{\mathrm{MLE}}).$$

Example (Fisher circle model, continued)

We have $g_{\omega\omega} = 1$ and $\overset{\mathrm{m}}{\Gamma}{}^{\omega}_{\omega\omega} = 0$.

Thus, the optimal orthogonal shift is

$$\begin{aligned} h_{\omega\omega}(x;\eta) &= \left(\partial_{\omega\omega}\eta_i - \overset{\mathrm{m}}{\Gamma}{}^{\omega}_{\omega\omega}\partial_\omega\eta_i\right)\partial_i p(x;\eta) \\ &= p(x;\eta)(-\cos\omega(x_1-\eta_1) - \sin\omega(x_2-\eta_2)) \end{aligned}$$

and the risk improvement by the optimal orthogonal shift is

$$\begin{aligned} &\mathrm{E}[D(p(y;\omega), p_{\alpha,0}(y;\hat{\omega}_{\mathrm{MLE}}))] - \mathrm{E}[D(p(y;\omega), p_{\alpha,\beta_{\mathrm{opt}}}(y;\hat{\omega}_{\mathrm{MLE}}))] \\ =&\frac{1}{8n^2}\overset{\mathrm{m}}{H}{}^{\lambda}_{ab}\overset{\mathrm{m}}{H}{}^{\kappa}_{cd}g^{ab}g^{cd}g_{\kappa\lambda} + \mathrm{o}(n^{-2}) \\ =&\frac{1}{8n^2}E[(-\cos\omega(y_1-\eta_1) - \sin\omega(y_2-\eta_2))^2] + \mathrm{o}(n^{-2}) \\ =&\frac{1}{8n^2} + \mathrm{o}(n^{-2}). \end{aligned}$$

The risk improvement by the optimal shift is $\frac{3}{8n^2} + o(n^{-2})$.

If the variance of x_1, x_2 is σ^2, the risk improvement by the extended plugin $p(y; \hat{\eta}_\pi)$ and by the Bayesian predictive density are

$$\frac{\sigma^2}{8n^2}, \quad \frac{\sigma^2 + 2}{8n^2},$$

respectively.

Therefore when σ^2 is large, the risk improvement by $p(y; \hat{\eta}_\pi)$ becomes relatively large, and the performance of the extended plugin density is close to that of the Bayesian predictive densityl.

From an information-geometric point of view, the orthogonal shift to $p(y; \hat{\eta}_\pi)$ is the projection of the orthogonal shift to the Bayesian predictive density onto the tangent space of $\mathscr{E}$, and the cosine of the angle between the two shift vectors is

$$\sqrt{\frac{\sigma^2}{8n^2}} \Big/ \sqrt{\frac{\sigma^2 + 2}{8n^2}} = \sqrt{\frac{\sigma^2}{\sigma^2 + 2}}\,.$$

Thus the angle between those shifts approaches to 0 as σ^2 grows.

Example 2-dimensional spike model

We consider the 2-dimensional spike model, namely 2-dimensional Gaussian distributions with zero mean vector and unknown covariance matrix that is expressed as

$$\Sigma = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} \lambda+1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix},$$

where $\lambda > 0$.

The eigenvalues of Σ are $\lambda + 1$ and 1.

The model $\mathscr{P} = \{N(0,\Sigma) \mid (\lambda,\phi)\}$ is parametrized by (λ,ϕ).

The components of Σ are the coordinate system (η_i) of the extended statistical model $\mathscr{E} = \{N(0,\Sigma) \mid \Sigma\}$, thus $p(y;\hat{\Sigma}_\pi)$ is the extended plugin density with the posterior mean of η.

In this model,

$$g_{\phi\phi} = \frac{\lambda^2}{\lambda+1},\ g_{\lambda\lambda} = \frac{1}{2(\lambda+1)^2},\ g_{\lambda\phi} = 0,$$

$$\overset{\mathrm{m}}{\Gamma}_{\phi\phi\lambda} = -\frac{\lambda}{(\lambda+1)^2},\ \overset{\mathrm{m}}{\Gamma}_{\phi\lambda\phi} = \frac{\lambda}{\lambda+1},\ \overset{\mathrm{m}}{\Gamma}_{\phi\phi\phi} = \overset{\mathrm{m}}{\Gamma}_{\phi\lambda\lambda} = \overset{\mathrm{m}}{\Gamma}_{\lambda\lambda\lambda} = \overset{\mathrm{m}}{\Gamma}_{\lambda\lambda\phi} = 0$$

hold.

The optimal orthogonal shift is

$$\begin{aligned}\frac{g^{ab}}{2n}h_{ab}(y;\eta) &= \frac{g^{ab}}{2n}\left(\partial_{ab}\eta_i - \overset{\mathrm{m}}{\Gamma}{}^{c}_{ab}\partial_c\eta_i\right)\partial_i p(y;\eta)\\ &= \frac{\lambda+1}{2n\lambda}\{(-y_1\sin\phi + y_2\cos\phi)^2 - 1\}p(y;\eta).\end{aligned}$$

The risk improvement by the optimal orthogonal shift is

$$\begin{aligned}
&\mathrm{E}[D(p(y;\omega), p_{\alpha,0}(y;\hat{\omega}_{\mathrm{MLE}}))] - \mathrm{E}[D(p(y;\omega), p_{\alpha,\beta_{\mathrm{opt}}}(y;\hat{\omega}_{\mathrm{MLE}}))] \\
=&\frac{1}{8n^2}\|2n\beta_{\mathrm{opt}}^{\kappa} h_{\kappa}\|^2 + \mathrm{o}(n^{-2}) \\
=&\frac{1}{8n^2} E\left[\left\{\frac{\lambda+1}{\lambda}((-y_1\sin\phi + y_2\cos\phi)^2 - 1)\right\}^2\right] + \mathrm{o}(n^{-2}) \\
=&\frac{1}{4n^2}\left(\frac{\lambda+1}{\lambda}\right)^2 + \mathrm{o}(n^{-2}).
\end{aligned}$$

The risk improvement by the optimal shift in $\mathscr{P}$ is

$$\frac{1}{4n^2}\left\{\left(\frac{\lambda+1}{\lambda}\right)^2 + 5\right\} + \mathrm{o}(n^{-2}).$$

Therefore, the risk improvement by the extended plugin grows as λ approaches to zero.

Much less numerical computation is required to otbain extended plugin densities than to obtain Bayesian predictive desities.

1. Introduction
2. Extended plugin densities
3. Information geometry of extended plugin densities
4. Shrinkage priors

We consider the choice of priors for the extended plugin with the Bayes estimator $\hat{\eta}_\pi$.
The Laplacian Δ on a manifold with the Riemannian metric g_{ab} is defined by

$$\Delta f := |g|^{-1/2} \partial_a (|g|^{1/2} g^{ab} \partial_b f),$$

where f is a smooth function on the model manifold.

A C^2 function f is called superharmonic if $\Delta f \leq 0$.

Theorem

Suppose that π is a smooth positive function on the model manifold $\mathscr{P}$. The extended plugin density $p(y; \hat{\eta}_\pi)$ based on π asymptotically dominates the extended plugin density $p(y; \hat{\eta}_{\pi_J})$ based on the Jeffreys prior π_J if and only if $(\pi/\pi_J)^{1/2}$ is a non-constant positive superharmonic function.

Proof.

The risk difference is

$$\begin{aligned}
&\mathrm{E}[D(p(y;u);p(y;\hat{\eta}_{\pi_J}))] - \mathrm{E}[D(p(y;u);p(y;\hat{\eta}_{\pi}))] \\
&= -\frac{2}{n^2}\left(\frac{\pi_J}{\pi}\right)^{1/2} \Delta \left(\frac{\pi}{\pi_J}\right)^{1/2} + \mathrm{o}(n^{-2}).
\end{aligned}$$

□

The superharmonic condition for priors for Bayesian predictive distributions is obtained in K(2006), and Theorem 4 is its parallel result for extended plugin distributions of Bayes estimators.

Shrinkage priors are closely related to the superharmonic condition.

For example, the Stein prior $\pi(\mu) = ||\mu||^{-(d-2)}$ for the estimation of mean vector μ of d-dimensional Gaussian distributions is and it satisfies the superharmonic condition when $d > 2$.

Consequently, Theorem suggests that shrinkage priors are effective for constructing an extended plugin of the Bayes estimator.

For the multivariate normal model with known covaraince matrix, finite sample theories have been developed, see K (2001), George et al. (2006), George et al. (2012).

Conclusion

1. The extended plugin density with the posterior mean of the expectation parameter η of the full exponential family is shown to be optimal regarding the Bayes risk based on the Kullback–Leibler divergence if we choose a predictive density from the full exponential family.
2. Several results are obtained from information-geometric viewpoints. The results are parallel to those for Bayesian predictive distributions.