10. *Robust Bayesian analysis under generalized moments conditions* (1994). With B.Betrò and M.Męczarski. Journal of Statistical Planning and Inference, vol .41, n. 2, 257-266.

11. *Robust Bayesian analysis given priors on partition sets* (1994). With C.Carota. To appear in Test, vol.3, n.2.

12. *Local and global sensitivity under some classes of priors* (1994). Recent Advances in Statistics and Probability, 233-243. J.Perez Vilaplana and M.L. Puri Editors. VSP, Ah Zeist.

13. *Density Based Classes of Priors: Infinitesimal Properties and Approximations* (1995). With L.Wasserman. To appear in Journal of Statistical Planning and Inference.

14. *On defining neighbourhoods of measures through the concentration functions* (1995). With S.Fortini. To appear in Sankhyā, Series A.

15. *Some results on posterior regret* Γ*−minimax estimation* (1995). With D.Ríos Insua and B.Vidakovic. Tentatively accepted by Statistics and Decision.

Editor. Springer-Verlag, Berlin.

2. *Posterior ranges of functions of parameters under priors with specified quantiles* (1990). Communications in Statistics - Theory and Methods, vol.19, n.1, 127-144.

3. *Robust Bayesian analysis given a lower bound on the probability of a set* (1991). Communications in Statistics - Theory and Methods, vol.20, n.5 & 6, 1881-1891.

4. *Conditional* Γ*–minimax actions under convex losses* (1992). With B. Betrò. Communications in Statistics - Theory and Methods, vol.21, n. 4, 1051-1066.

5. *Bounds on the prior probability of a set and robust Bayesian analysis* (1993). Theory of Probability and Its Applications, Vol.37, n.2, 358-359.

6. *Infinitesimal Sensitivity of Posterior Distributions* (1993). With L. Wasserman. Canadian Journal of Statistics, vol.21, n.2, 195-203.

7. *Robust Bayesian analysis given bounds on the probability of a set* (1993). Communications in Statistics - Theory and Methods, vol.22, n.11, 2983-2998.

8. *Concentration function and coefficients of divergence for signed measures* (1993). With S.Fortini. Journal of the Italian Statistical Society, vol. 2, n.1, 17-34.

9. *Concentration functions and Bayesian robustness* (1994). With S.Fortini. Journal of Statistical Planning and Inference, vol. 40, n.2 & 3, 205-220.

Biography

Fabrizio Ruggeri was born on May 1, 1956 in Reggio nell'Emilia, Italy. He obtained a Bachelor's degree in Mathematics from Università degli Studi di Milano, Italy, in 1982 and a Master's degree in Statistics from Carnegie Mellon University, Pittsburgh, Pennsylvania in 1989. In 1983 he worked as a high school teacher in Milano, Italy; from 1983 to 1985 he worked as an applied Mathematician, expert of optimisation techniques, at the Research and Development Division of Alfa Romeo Auto Spa, Arese (Milano), Italy; from 1985 to 1987, he worked as a Computer Consultant at Gaetani Spa, Milano, Italy; from 1987 to 1988 he got two fellowships from Consiglio Nazionale delle Ricerche (CNR), the former to work in Milano, at the Istituto per le Applicazioni della Matematica e dell'Informatica (IAMI) of CNR, and the latter to study in Pittsburgh, at the Department of Statistics of Carnegie Mellon University; since 1988 he holds a permanent position as a Researcher at CNR-IAMI. He is the author of some papers, including the following:

1. *Measures of dissimilarities for contrasting information sources in data fusion* (1989). With L.Olivi and R.Rotondi. Reliability Data Collection and Use in Risk and Availability Assessment. (Proceedings of the 6^{th} EuReData Conference, Siena, March 15-17, 1989), 671-681. V. Colombari

[68] Wasserman, L.A. and Seidenfeld, T., *The dilation phenomenon in robust Bayesian inference*, Journal of Statistical Planning and Inference, 1994, Vol. 40, pages 345-356.

[59] Ruggeri, F., *Robust Bayesian analysis given bounds on the probability of a set*, Communications in Statistics - Theory and Methods, 1993, Vol.22, pages 2983-2998.

[60] Ruggeri, F., *Local and global sensitivity under some classes of priors*, in New Progress in Probability and Statistics (Vilaplana, J.P. and Puri, M.L. Eds.), VSP, Ah Zeist, 1994, pages 233-243.

[61] Ruggeri, F., *Discussion of "An overview of robust Bayesian analysis"*, Test, 1994, Vol. 3, pages 108-109.

[62] Ruggeri, F. and Wasserman L.A., *Infinitesimal sensitivity of posterior distributions*, Canadian Journal of Statistics, 1993, Vol. 21, pages 195-203.

[63] Ruggeri, F. and Wasserman L.A., *Density based classes of priors: infinitesimal properties and approximations*, to appear in Journal of Statistical Planning and Inference, 1995.

[64] Seidenfeld, T. and Wasserman, L.A., *Dilation for sets of probabilities*, Annals of Statistics, 1993, Vol. 21, pages 1139-1154.

[65] Vajda, I., *Distances and discrimination rates for stochastic processes*, Stochastic Processes and their Applications, 1990, Vol. 35, pages 47-57.

[66] Yamato, H., *Characteristic functions of means of distributions chosen from a Dirichlet process*, Annals of Probability, 1984, Vol. 12, pages 262-267.

[67] Wasserman, L.A., *Recent methodological advances in robust Bayesian inference*, in Bayesian Statistics IV (Bernardo, J., Berger, J., Dawid, A.P. and Smith, A.F.M. Eds.) Oxford Press, Oxford, 1992, pages 483-502.

[51] Ríos Insua, D., Ruggeri, F. and Vidakovic, B., *Some results on posterior regret* Γ*–minimax estimation*, tentatively accepted by Statistics and Decision, 1995.

[52] Rizzi, A., *Osservazioni sulle classi di Fréchet a più variabili*, Bollettino dell'Unione Matematica Italiana, 1957, Vol. 12, pages 263-277.

[53] Rohatgi, V.K., *Statistical Inference*, John Wiley and Sons, New York, 1984.

[54] Rotondi, R., Ruggeri, F. and Vercesi P., *A Bayesian approach to comparison of information sources*, Istituto per le Applicazioni della Matematica e dell'Informatica, Consiglio Nazionale delle Ricerche, Milano, Italy, 1992, Quaderno 92.10.

[55] Ruggeri, F., *Posterior ranges of functions of parameters under priors with specified quantiles*, Communications in Statistics - Theory and Methods, 1990, Vol.19, pages 127-144.

[56] Ruggeri, F., *Robust Bayesian analysis given a lower bound on the probability of a set*, Communications in Statistics - Theory and Methods, 1991, Vol.20, pages 1881-1891.

[57] Ruggeri, F., *Bayesian comparison of Italian earthquakes*, Institute of Statistics and Decision Sciences, Duke University, Durham, 1993, Discussion Paper 93-A32.

[58] Ruggeri, F., *Bounds on the prior probability of a set and robust Bayesian analysis*, Theory of Probability and Its Applications, 1993, Vol.37, pages 358-359.

[42] Gini, C., *Sulla misura della concentrazione della variabilità dei caratteri*, Atti del Reale Istituto Veneto di Scienze, Lettere ed Arte, Anno Accademico 1913-1914, 1914, Vol. 73, parte II, pages 1203-1248.

[43] Hollander, M. and Wolfe, D.A., *Nonparametric Statistical Methods*, John Wiley and Sons, New York, 1973.

[44] Karnik, V., *The seismicity of the European Area*, Hingham, London, 1971.

[45] Kendall, M.G., *A new measure of rank correlation*, Biometrika, 1938, Vol. 30, pages 81-93.

[46] Ladelli, L., Mitrione, M., Ruffoni, S. and Ruggeri, F., *Inferences for point processes modelling earthquake occurrences in Sannio Matese*, Istituto per le Applicazioni della Matematica e dell'Informatica, Consiglio Nazionale delle Ricerche, Milano, Italy, 1992, Quaderno 92.7.

[47] Lavine, M., *Some aspects of Polya tree distributions for statistical modelling*, Annals of Statistics, 1992, Vol. 20, pages 1222-1235.

[48] Marshall, A.W. and Olkin, I., *Inequalities: theory of majorization and its applications*, Academic Press, New York, 1979.

[49] Pietra, G. (1915), *Delle relazioni tra gli indici di variabilità*, Atti del Reale Istituto Veneto di Scienze, Lettere ed Arte, Anno Accademico 1914-1915, 1915, Vol. 74, parte II, pages 775-792.

[50] Regazzini, E., *Concentration comparisons between probability measures*, Sankhyā, Series B, 1992, Vol. 54, pages 129-149.

Basu (Ghosh M. and Pathak P.K. Eds.), Lecture Notes IMS, 1992, Vol. 17, pages 127-150.

[34] Feron, R., *Sur les tableaux de corrélation dont les marges sont données, cas de l'éspace à trois dimensions*, Publication de l'Institute de Statistique de l'Université de Paris, 1956, Vol. 12, pages 103-116.

[35] Fortini, S. and Ruggeri, F., *Concentration Function and Sensitivity to the Prior*, Institute of Statistics and Decision Sciences, Duke University, Durham, 1993, Discussion Paper 93-A33.

[36] Fortini, S. and Ruggeri, F., *Concentration function and coefficients of divergence for signed measures*, Journal of the Italian Statistical Society, 1993, Vol. 2, pages 17-34.

[37] Fortini, S. and Ruggeri, F., *Concentration functions and Bayesian robustness*, Journal of Statistical Planning and Inference, 1994, Vol. 40, pages 205-220.

[38] Fortini, S. and Ruggeri, F., *On defining neighbourhoods of measures through the concentration function*, to appear in Sankhyā, Series A, 1995.

[39] Fréchet M., *Sur les tables de corrélation dont les marges sont données*, Annales de l'Universite de Lyon, Sciences, 1951, Vol. 4, pages 13-84.

[40] Fréchet M., *Les tableaux de corrélation et les programmes linéaires*, Revue Institute Internationale de Statistique, 1957, Vol. 25, pages 23-40.

[41] Genest, C. and MacKay, R.J., *Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données*, Canadian Journal of Statistics, 1986, Vol. 14, pages 145-159.

Salinetti, G. Eds.), Kluwer Academic Publishers, Dordrecht, 1991, pages 1-12.

[25] Dall'Aglio, G., Kotz, S. and Salinetti, G. (Eds.) *Advances in Probability Distributions with Given Marginals*, Kluwer Academic Publishers, Dordrecht, 1991.

[26] Zen, M. and DasGupta, A., *Estimating a binomial parameter: is robust Bayes real Bayes?*, Statistics and Decisions, 1993, Vol. 11, pages 37-60.

[27] Diaconis, P. and Freedman, D., *On the consistency of Bayes estimates*, Annals of Statistics, 1986, Vol. 14, pages 1-67.

[28] Doksum, K., *Tailfree and neutral random probabilities and their posterior distributions*, Annals of Probability, 1974, Vol. 2, pages 183-201.

[29] Doss, H. and Sellke, T., *The tails of probabilities chosen from a Dirichlet prior*, Annals of Statistics, 1982, Vol. 10, pages 1302-1305.

[30] Ferguson, T.S., *A Bayesian Analysis of some Nonparametric Problems*, Annals of Statistics, 1973, Vol. 1, pages 209-230.

[31] Ferguson, T.S., *Prior Distribution on Spaces of Probability Measures*, Annals of Statistics, 1974, Vol. 2, pages 615-629.

[32] Ferguson, T.S. and Phadia, E.G., *Bayesian nonparametric estimation based on censored data*, Annals of Statistics, 1979, Vol. 7, pages 163-186.

[33] Ferguson, T.S., Phadia, E.G. and Tiwari, R.C. *Bayesian Nonpametric Inference*, in Current Issues in Statistical Inference: Essays in honor of D.

[16] Cifarelli, D.M. and Regazzini, E., *On a general definition of concentration function*, Sankhyā, Series B, 1987, Vol. 49, pages 307-319.

[17] Cifarelli, D.M. and Regazzini, E., *Distribution functions of means of a Dirichlet process*, Annals of Statistics, 1990, Vol. 18, pages 429-442.

[18] Cifarelli, D.M. and Regazzini, E., *Correction to "Distribution functions of means of a Dirichlet process"*, to appear in Annals of Statistics, 1994, Vol. 22.

[19] Çinlar, E., *Introduction to Stochastic Processes*, Prentice-Hall, Englewood Cliffs, 1975.

[20] Csiszár, I., *Information-type measures of difference of probability distributions and indirect observations*, Studia Scientiarum Mathematicarum Hungarica, 1967, Vol. 2, pages 299-318.

[21] Dalal, S.R. and Phadia, E.G., *Nonparametric Bayes Inference for Concordance in Bivariate Distributions*, Communications in Statistics - Theory and Methods, 1983, Vol. 12, pages 947-963.

[22] Dall'Aglio, G., *Sugli estremi dei momenti delle funzioni di ripartizione doppia*, Annali delle Scuola Normale Superiore di Pisa, Classe di Scienze, 1956, Vol. 10, pages 35-74.

[23] Dall'Aglio, G., *Fréchet classes and compatibility of distribution functions*, Symposia Mathematica, 1972, Vol.9 , pages 131-150.

[24] Dall'Aglio, G., *Fréchet classes: the beginnings*, in Advances in Probability Distributions with Given Marginals, (Dall'Aglio, G., Kotz, S. and

[7] Berger, J., *An overview of robust Bayesian analysis (with discussion)*, Test, 1994, Vol. 3, pages 5-124.

[8] Bertino, S., *Su di una sottoclasse della classe di Fréchet*, Statistica, 1968, Vol. 25, pages 511-542.

[9] Bertino, S., *Sulla distanza fra distribuzioni*, Pubblicazioni dell'Istituto di Calcolo delle Probabilità dell'Università di Roma, 1968, Nr. 82.

[10] Betrò, B., Garavaglia, E., Guagenti, E., Rotondi, R. and Tagliani, A., *Sulla distribuzione dei tempi di intercorrenza fra eventi sismici in alcune zone italiane*, Atti *IV* Convegno Nazionale di Ingegneria Sismica, Milano, 1989.

[11] Betrò, B., Męczarski, M. and Ruggeri, F., *Robust Bayesian analysis under generalized moments conditions*, Journal of Statistical Planning and Inference, 1994, Vol .41, pages 257-266.

[12] Betrò, B. and Ruggeri, F., *Conditional* $\Gamma-$*minimax actions under convex losses*, Communications in Statistics - Theory and Methods, 1992, Vol.21, pages 1051-1066.

[13] Billingsley, P., *Probability and Measure*, John Wiley and Sons, New York, 1986.

[14] Campbell, K. W., *Bayesian analysis of extreme earthquake occurrences. Part I. Probabilistic hazard model*, Bulletin of the Seismological Society of America, 1982, Vol. 12, pages 1689-1705.

[15] Carota, C. and Ruggeri, F., *Robust Bayesian analysis given priors on partition sets*, to appear in Test, 1994, Vol.3.

Bibliography

[1] Abramowitz, M. and Stegun I.A., *Handbook of Mathematical Functions*, Dover, New York, 1972.

[2] Ali, S.M. and Silvey, S.D., *A general class of coefficients of divergence of one distribution from another*, Journal of the Royal Statistical Society, Series B, 1966, Vol. 28, pages 131-142.

[3] Antoniak, C.E., *Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems*, Annals of Statistics, 1974, Vol. 2, pages 1152-1174.

[4] Berger, J., *The robust Bayesian viewpoint (with discussion)*, in Robustness of Bayesian Analysis, (J.Kadane, Ed.), North Holland, Amsterdam, 1984, pages 63-124.

[5] Berger, J., *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York, 1985.

[6] Berger, J., *Robust Bayesian Analysis : sensitivity to the prior*, Journal of Statistical Planning and Inference, 1990, Vol. 25, pages 303-328.

Definition D.1.3 *B dilates (strictly dilates) A if* $[\underline{P}(A), \overline{P}(A)] \subseteq (\subset)$ $[\underline{P}(A|B), \overline{P}(A|B)]$.

Definition D.1.4 *The finite partition $\mathcal{B}$, for which $\underline{P}(B) > 0$ for all $B \in \mathcal{B}$, dilates (strictly dilates) A if every $B \in \mathcal{B}$ dilates (strictly dilates) A.*

Definition D.1.5 *$\mathcal{M}$ is dilation prone if there exists A and $\mathcal{B}$ such that $\mathcal{B}$ dilates A. Otherwise, $\mathcal{B}$ is dilation immune.*

Appendix D

Dilation

Given a class of probability measures, dilation occurs if there exists a subset A and a partition $\mathcal{B}$ such that the interval determined by the probability $P(A)$ is contained (in the set-theoretic sense) in the one determined by the conditional probability $P(A|B)$, for all $B \in \mathcal{B}$. Dilation is studied in the papers by Seidenfeld and Wasserman [64], [68]; we present here some definitions contained therein.

Let $\mathcal{X}$ be a nonempty set and $\mathcal{A}$ an algebra of subsets of $\mathcal{X}$. Let $\mathcal{M}$ be a subset of the space $\mathcal{P}$ of all probability measures on $(\mathcal{X}, \mathcal{A})$.

Definition D.1.1 *Upper probability function $\overline{P}$ and lower probability function $\underline{P}$ are given by*

$$\overline{P}(A) = \sup_{P \in \mathcal{M}} P(A) \text{ and } \underline{P}(A) = \inf_{P \in \mathcal{M}} P(A).$$

Definition D.1.2 *If $\underline{P}(B) > 0$, then conditional upper and lower probability given B are, respectively,*

$$\overline{P}(A|B) = \sup_{P \in \mathcal{M}} \frac{P(A \cap B)}{P(B)} \text{ and } \underline{P}(A|B) = \inf_{P \in \mathcal{M}} \frac{P(A \cap B)}{P(B)}.$$

$$\lambda(x, \mu; \tau, B) = \frac{1}{\pi}\int_{\tau}^{x} |x-y|^{\beta+1} \exp(-\int_{\tau}^{\infty} \log|u - i\mu - y| dB(u)) \sin(\pi B(y)) dy.$$

The previous Theorem can be applied in order to determine the pdf of

$$Y_\psi = \int_{\Re} \psi(x) P(dx),$$

where ψ is a measurable function such that the following condition holds.

Condition C.1.2.

$$\mathcal{P}\left\{P : \int_{\Re} |\psi| dP < \infty\right\} = 1.$$

Under these conditions, $P\psi^{-1}(\cdot) := P(\psi^{-1}(\cdot))$ turns out to be a random probability measure chosen by the Dirichlet process on $(\Re, \mathcal{B})$ with parameter $\eta_\psi = \eta(\psi^{-1}(\cdot))$.

Corollary C.1.1 (Cifarelli and Regazzini, [17] and [18]). *Let P be a random probability measure chosen by a Dirichlet process on $\Re, \mathcal{B}$, with parameter η; let $\psi : \Re \to \Re$ be a measurable function satisfying Condition C.1.2. and let $\mathcal{M}_\psi$ be the pdf of Y_ψ. Under these conditions, $\mathcal{M}_\psi$ coincides with $\mathcal{M}$ in Theorem C.1.1 upon replacement of B by the df corresponding to η_ψ.*

we obtain

(i) for $\inf S(\eta) = \tau > -\infty$ *and* $B(\tau) \geq 1$*, it follows that* $\mathcal{M}(x) =$,

$$\begin{cases} 0 & x < \tau \\ \int_\tau^x \frac{2^{\beta-3}(\beta-1)}{\pi(u-\tau)} du \int_\pi^\pi \{\cos(y/2)\}^{\beta-2} \cdot & \\ \cdot \cos\left\{\int_\tau^\infty q(v;u,y)(u-\tau)\sin y dv - \frac{\beta y}{2}\right\} \cdot & \\ \cdot \exp\{-\int_\tau^\infty q(v;u,y)[(u-\tau)\cos y + v - \tau]dv\}dy & x \geq \tau, \end{cases}$$

where

$$q(v;u,y) = \frac{\beta - B(v)}{(u-\tau)^2 + (v-\tau)^2 + 2(v-\tau)(u-\tau)\cos y};$$

(ii) for $\inf S(\eta) = \tau > -\infty$ *and* $B(\tau) \in [0,1)$,

$$\mathcal{M}(x) = l(x;\tau,B) = \lim_{\mu\to 0^+} \lambda(x,\mu;\tau,B) \quad x \in (\tau, \sup S(\eta));$$

(iii) for $\inf S(\eta) = -\infty$ *and* $x < \sup S(\eta)$,

$$\mathcal{M}(x) = \lim_{\tau\to-\infty} l(x;\tau,B_\tau) = \lim_{\tau\to-\infty} \lim_{\mu\to 0^+} \lambda(x,\mu;\tau,B_\tau) x \in (\tau, \sup S(\eta)),$$

where

$$B_\tau(u) = \begin{cases} 0 & u < \tau \\ B(u) & u \geq \tau, \end{cases}$$

$$\begin{aligned} l(x;\tau,B) \quad &= \frac{2^{\beta-1}}{\pi}(x-\tau)^\beta \int_0^\pi (\cos y/2)^{\beta-1} \exp\{-\int_\tau^\infty \log|(x-\tau)e^{iy} + \\ &u - \tau|dB(u)\}\cos(\frac{\beta+1}{2}y - \int_\tau^\infty \arg((x-\tau)e^{iy}u - \tau)dB(u))dy, \end{aligned}$$

Appendix C

Distribution of random functionals

We present now some results due to Cifarelli and Regazzini ([17], [18]), mentioned in Chapter 5 on nonparametric Bayesian robustness, on the distribution of random functionals depending on a random probability measure P on $(\Re, \mathcal{B})$ chosen from a Dirichlet process.

We need first some notations and a condition ensuring that each random probability measure P chosen from a Dirichlet process with parameter $\eta = \beta\alpha$, $\alpha(\Re) = 1$, is a probability distribution function (pdf) with finite expectation.

Let $\mathcal{M}$ be the pdf of $Y = \int_{\Re} x dP$, $S(\eta)$ be the support of the parameter η, $\Upsilon(x) = P((-\infty, x])$ and $B(x) = \beta\alpha((-\infty, x])$.

Condition C.1.1

$$\mathcal{P}\left\{\Upsilon : \Upsilon \text{ is a pdf and } \int_{\Re} |x| d\Upsilon < \infty\right\} = 1$$

Theorem C.1.1 (Cifarelli and Regazzini, [17] and [18]). *Let P be a random probability measure chosen by a Dirichlet process on $(\Re, \mathcal{B})$, with parameter η, and satisfying Condition C.1.1. Then, if η is degenerate at μ, $\mathcal{M}$ is also degenerate at the same point. On the other hand, if η is not degenerate,* A total ordering, consistent with the previous partial one, is achieved when considering $C(\Pi)$ and $G(\Pi)$, as shown by Cifarelli and Regazzini [16]; in general, the coefficients of divergence allows for such a total ordering, as shown in Chapter 5 when using the Kullback-Leibler index to define the distance between Dirichlet processes.

For further results on the c.f., we refer, besides to the above mentioned papers by Cifarelli and Regazzini [16] and Regazzini [50], to the subsequent ones, mainly applied to Bayesian robustness, due to Fortini and Ruggeri [35], [36], [37] and [38], Ruggeri [60], Rotondi, Ruggeri and Vercesi [54] and Carota and Ruggeri [15].

divergence. We present now a partial ordering, induced by the c.f., in the space $\mathcal{P}$ of all probability measures.

Definition B.1.2 *Let φ_1, φ_2 denote the c.f.'s of Π_1 and Π_2 w.r.t. Π_0; if $\varphi_2(x) \leq \varphi_1(x) \quad, \forall x \in [0,1]$, we will say that Π_2 is not less concentrated than Π_1 w.r.t. Π_0 and we write $\Pi_2 \succeq \Pi_1$.*

As stated in the next Theorem, due to Regazzini [50], the previous partial ordering is equivalent to the one induced by the class of coefficients of divergence, considered in Ali and Silvey [2] and Csiszár [20],

$$\rho(\Pi, g) = \int_{[0,\infty)} g(t) dH_\Pi(t) + \Pi_s(\Theta) \lim_{t\to\infty} \{g(t)/t\},$$

where $g : [0,\infty) \to \Re$ is continuous and convex, whereas H_Π and Π_s are defined as before, for any $\Pi \in \mathcal{P}$, with respect to a fixed $\Pi_0 \in \mathcal{P}$.

Theorem B.1.2 *For any pair of probability measures $\Pi_1, \Pi_2 \in \mathcal{P}$, $\Pi_1 \preceq \Pi_2$ holds w.r.t. Π_0 if and only if $\rho(\Pi_1, g) \leq \rho(\Pi_2, g)$ for all continuous, convex g for which $\rho(\Pi_1, g)$ and $\rho(\Pi_2, g)$ are finite.*

The Gini's concentration ratio (Gini [42]) $C(\Pi) = 2\int_0^1 \{x - \varphi(x)\}\, dx$ and the index $G(\Pi) = \sup_{x\in[0,1]} \{x - \varphi(x)\}$ proposed by Pietra [49], which equals twice the variational distance $\sup_{A\in\mathcal{F}} |\Pi(A) - \Pi_0(A)|$, are obtained as particular cases of $\rho(\Pi, g)$, taking, respectively,

$$g(t) = 1/2 \int_\Re |t - u| dH_\Pi(u) \; + \; 1/2 \; \Pi_S(\Theta) \text{ and } g(t) = |t-1|.$$

Besides, when Π is absolutely continuous w.r.t Π_0, the Kullback-Leibler index and the χ^2 divergence are obtained from $\rho(\Pi, g)$ setting $g(t) = t\log t$ and $g(t) = (t-1)^2$.

sure and a finite, positive one, with equal total mass (as studied by Fortini and Ruggeri [36]). Here, we suggest its extension to comparison of finite, positive measures Π and Π_0 with different total mass, say β and β_0, respectively. Obvious changes occur in the definition of c.f. and in its properties, e.g. $\varphi(\beta_0) = \Pi_a(\Theta)$ and $\varphi(x) = \dfrac{\beta}{\beta_0} x \;\forall x \in [0, \beta_0] \Leftrightarrow \Pi = \dfrac{\beta}{\beta_0}\Pi_0$.

The following Theorem, proved in Cifarelli and Regazzini [16], is very useful in understanding the meaning of the c.f.; it states that $\varphi(x)$ substantially coincides with the minimum value of Π on the measurable subsets of Θ with Π_0-measure not smaller than x.

Theorem B.1.1 *If $A \in \mathcal{F}$, $\Pi_0(A) = x$, then $\varphi(x) \leq \Pi_a(A)$. Moreover if $x \in [0, 1]$ is adherent to the range of H, then B_x exists such that $\Pi_0(B_x) = x$ and*

$$\varphi(x) = \Pi_a(B_x) = min\{\Pi(A) : A \in \mathcal{F} \text{ and } \Pi_0(A) \geq x\}. \tag{B.1}$$

If Π_0 is nonatomic, then (B.1) holds for any $x \in [0, 1]$.

Such a Theorem is relevant in applying the c.f. to robust Bayesian analysis; in fact, given any $x \in [0, 1]$, the probability, under Π, of all the subsets A with Π_0–measure x, is such that $\varphi(x) \leq \Pi(A) \leq 1 - \varphi(1 - x)$.

As described in Fortini and Ruggeri [35], [37] and [38] and Ruggeri [60], the c.f. can be used to define neighbourhoods Γ of a baseline probability measures and compute bounds on posterior expectations when the measures vary in Γ; besides, the c.f. can be used to check if the posterior measures, obtained by updating the priors in a class Γ, are in a neighbourhood of a baseline posterior one.

In Chapter 5, we compare probability measures by introducing an ordering on them based on the c.f. and some related indices, known as coefficients of

$\Pi_a(\cdot) = \int_{\cdot \cap N^C} h(\theta)\Pi_0(d\theta)$ and Π_s denote the absolutely continuous and the singular part of Π w.r.t. Π_0, respectively. Set $h(\theta) = \infty$ all over N and define $H(y) = \Pi_0(\{\theta \in \Theta : h(\theta) \leq y\}), c_x = \inf\{y \in R : H(y) \geq x\}$. Finally, let $L_x = \{\theta \in \Theta : h(\theta) \leq c_x\}$ and $L_x^- = \{\theta \in \Theta : h(\theta) < c_x\}$.

Definition B.1.1 *The function $\varphi : [0,1] \to [0,1]$ is said to be the concentration function of Π w.r.t. Π_0 if $\varphi(x) = \Pi(L_x^-) + c_x\{x - H(c_x^-)\}$ for $x \in (0,1)$, $\varphi(0) = 0$ and $\varphi(1) = \Pi_a(\Theta)$.*

Observe that $\varphi(x)$ is a nondecreasing, continuous and convex function, such that $\varphi(x) \equiv 0 \Leftrightarrow \Pi \perp \Pi_0$, $\varphi(x) = x \; \forall x \in [0,1] \Leftrightarrow \Pi = \Pi_0$ and

$$\varphi(x) = \int_0^{c_x} [x - H(t)]dt = \int_0^x c_t dt.$$

Furthermore, $\varphi(1) = 1$ characterises the measures Π's which are absolutely continuous w.r.t. Π_0, whereas $\varphi(x) = 0, 0 \leq x \leq \alpha$, means that Π gives no mass to a subset $A \in \mathcal{F}$ such that $\Pi_0(A) = \alpha$. To show how to practically draw the c.f. when Π_0 and Π are absolutely continuous w.r.t. Lebesgue measure, consider, as an example, $\Theta \equiv \Re^+$, the Borel σ–algebra $\mathcal{F}$ on it, $\Pi \sim \mathcal{G}(2,2)$ and $\Pi_0 \sim \mathcal{E}(1)$; then it follows that the Radon-Nikodym derivative is $h(\theta) = 4\theta^2 e^{-\theta}$. The c.f. $\varphi(x)$ is obtained by evaluating $x = \Pi_0(L_q)$ and $\varphi(x) = \Pi(L_q)$, where $L_q = \{\theta \in \Theta : h(\theta) \leq q\}$ and q takes as many values as possible to draw the c.f. within the required accuracy.

Note that the c.f. is not symmetric, i.e. in general the c.f. φ_1 of Π w.r.t. Π_0 is different from the c.f. φ_2 of Π_0 w.r.t. Π, since it can be easily shown that the two pairs $(y, \varphi_2(y))$ and $(1 - \varphi_1(1-x), x)$ coincide.

The c.f. can be defined even between finite, positive measures with equal total mass (as stated in Cifarelli and Regazzini [16]), between a signed mea-

Appendix B

Concentration function and coefficients of divergence

The concentration function and the coefficients of divergences are tools which are used in statistical analysis to compare probability measures and measure their distance. In this appendix, we present their main properties which are used in Chapter 5 on nonparametric robustness.

Cifarelli and Regazzini [16] defined the concentration function (c.f.), as a generalisation of the Lorenz curve, described in Marshall and Olkin [48], pag.5. The classical definition of concentration refers to the discrepancy between a discrete probability measure Π and a uniform one, say Π_0, and allows for the comparison of the two probability measures, looking for subsets where Π is much more concentrated than Π_0 (and vice versa). Cifarelli and Regazzini [16] defined and studied the c.f. of Π with respect to (w.r.t.) Π_0, where Π and Π_0 are two probability measures on the same measurable space $(\Theta, \mathcal{F})$. According to the Radon-Nikodym theorem, there is a partition $\{N, N^C\} \subset \mathcal{F}$ of Θ and a nonnegative function h on N^C such that, $\forall E \in \mathcal{F}$, $\Pi(E) = \int_{E \cap N^C} h(\theta)\Pi_0(d\theta) + \Pi_s(E \cap N)$, $\Pi_0(N) = 0$, $\Pi_s(N) = \Pi_s(\Theta)$, where *that* $0 \leq f_n \uparrow f$ *almost everywhere, then* $\int f_n d\mu \uparrow \int f d\mu$.

Theorem A.1.3 (Lebesgue's dominated convergence) *If an integrable function* g *exists such that* $|f_n| \leq g$ *almost everywhere, and if* $f_n \to f$ *almost everywhere, then* f *and the* f_n *are integrable and* $\int f_n d\mu \to \int f d\mu$.

Appendix A

Basic probability theorems

The next definition and results are well known and can be found in many probability books, e.g. Billingsley [13], p.275.

Definition A.1.1 *Given the random variables* $X_1, \ldots, X_n$ *on the probability space* $(\Omega, \mathcal{F}, P)$*, the empirical distribution function for* $X_1, \ldots, X_n$ *is the distribution function* $F_n(x, \omega)$*,* $\omega \in \Omega$*, with a jump of* $1/n$ *at each* $X_k(\omega)$*.*

If the random variables X_k are independent and identically distributed (i.i.d.), with distribution $F(x)$, then $F_n(x, \omega)$ is its natural estimate; the limiting behaviour of $F_n(x, \omega)$ is given by the Glivenko-Cantelli theorem.

Theorem A.1.1 *Suppose that* $X_1, X_2, \ldots$ *are i.i.d. random variables with common distribution function* F*, then* $\sup_x |F_n(x, \omega) - F(x)| \to 0$ *with probability one.*

We need also some convergence theorem for integrals of measurable functions f on the measurable space $(\Omega, \mathcal{F})$, with respect to a measure μ (see Billingsley [13], p. 211-213). In the next, we will assume that all functions are measurable.

Theorem A.1.2 (Monotone convergence) *Given a sequence* $\{f_n\}$ *such* applied to them as well.

Finally, another important, open problem is represented by modeling data with given marginals, in a way that the nonparametric Bayesian approach preserves them. That should lead to consider mixtures of Dirichlet processes, like in Antoniak [3].

Chapter 7

Conclusions

We have mainly considered connections between two important areas in the statistical research, nonparametric Bayesian inference and Fréchet classes, but we have also proved some results which are more pertinent to only one of them (e.g. the Bayesian interpretation of the Kolmogorov-Smirnov test and the properties of the updated classes of probabilities $\Pi(\hat{W}, \hat{M})$ and $\Psi(\hat{W}, \hat{M})$).

We want to mention here only few points which represents, in our opinion, the most important areas in which work is to be done in the very next future.

During this research, we came out with a new idea about nonparametric Bayesian robustness: its extension to classes different from the Fréchet one started here, but other classes can be studied too (e.g. density bounded, density ratio). In general, also the other problems studied here could be faced by considering a class of Dirichlet processes with parameters in other classes of measures. Another way to define the distance should be pursued: the Prohorov distance between spaces of probability measures.

In literature, many subclasses of the Fréchet class have been studied (see e.g. Genest and MacKay [41]) and the ideas contained in this work could be

6.2.4 Future developments

We mention here a couple of possible, future developments of the analysis: distance from a given distribution, e.g. the independent one with marginals distributions $F(x)$ and $G(y)$ (shortly discussed in Section 4.6) and prediction. We could suppose that a certain time x_0, expressed in years, passed since the last major earthquake and that y is the number of minor earthquakes occurred since then. We could then be interested in the predictive probability for X and upper and lower bounds on it (see Section 4.4. for more details).

Table 6.3: Distance between updated Fréchet bounds

β	Distance
1	.0125805340
5	.0571842454
10	.1026982367
20	.1705835118
30	.2187918956
40	.2547956252
50	.2827086291
60	.3049826423
70	.3231696806
80	.3383000738
90	.3510846696
100	.3620297552

achieve its maximum at $(2.11423517, 59.43643323)$ where F and G coincide and equal $1/2$. Following a reasoning similar to the one used in proving Theorem 4.1.5, it can be seen that the distance achieves its maximum at one of the points in which either $Y = 59$ or $Y = 60$ and either $F(x) = G(y)$ or $F(x) + G(y) = 1$. In such points, it turns out that

$$\begin{aligned}\sup_{(x,y)\in\Re^2}\{M(x,y) - W(x,y)\} &= \max\{G(60), G(59), 1 - G(60), 1 - G(59)\}\\ &= 0.5032213598.\end{aligned}$$

If we suppose that $\beta = 1$, then it follows that

$$\sup_{(x,y)\in\Re^2}\{\hat{M}(x,y) - \hat{W}(x,y)\} = 0.5032213598/40 = .01258053400.$$

Therefore, in this case the maximum distance between upper and lower Bayes estimators is very small, being they strongly influenced by data, much more than by prior knowledge. Such a conclusion holds even for larger values of β, as shown by Table 6.3.

Table 6.2: Bounds on concordance

β	Lower bound	Upper bound
1	.7989331880	.8146929475
5	.7471259760	.8224367545
10	.6901070255	.8321249690
20	.5972027845	.8499544885
30	.5254672765	.8650927865
40	.4687483355	.8777506305
50	.4229145395	.8883640910
60	.3851691804	.8973368990
70	.3535761264	.9049958965
80	.3267614895	.9115961370
90	.3037271217	.9173352075
100	.2837321190	.9223667315

$\beta = 30$, there is concordance between X and Y: this is not unexpected, since we are quite confident that if one observation has greater interoccurrence time than another, then also its number of minor earthquakes should be greater.

6.2.3 Decision

The Bayes estimators of the expected values of X and Y, under squared loss function, discussed in Section 4.1.2, are 3.050197415 and 86.71730745, respectively.

Consider now the Bayes estimators of the random distribution function under squared loss function; from Ferguson [30], we conclude that the class of such estimators coincides with $\Pi(\hat{W}, \hat{M})$. Comparing such lower and upper bounds could be interesting and we consider the distance $\sup_{(x,y)\in\Re^2} \{\hat{M}(x,y) - \hat{W}(x,y)\}$.

Since Y is discrete, we cannot apply Theorem 4.1.5 but a slightly modified result. Consider Y as if it were not discrete, then $M(x,y) - W(x,y)$ would

$$W(x,y) = \max\left\{1 - e^{-0.3278477195x} - 0.98859649^{y+1}, 0\right\}.$$

6.2.2 Concordance

Using the same notations as in Section 2.3.1, we can compute upper and lower bounds on the Bayes estimator of the concordance coefficient Δ_H under the squared error loss function, as given by Theorem 4.2.2.

Since $P_\alpha\{X = X'\} = 0$, $P_M\{(X-X')(Y-Y') \geq 0\} = 1$, $P_W\{(X-X')(Y-Y') \leq 0\} = 1$, and $P_\alpha\{Y = Y'\} = \sum_{k=1}^{\infty} p_k^2 = \dfrac{p}{2-p}$, where $p_k = P_\alpha\{Y = k\}$, it follows that

$$\sup \Delta_\alpha = 1 - P_\alpha\{Y = Y'\}/2 = \frac{4-3p}{4-2p}, \text{ and}$$
$$\inf \Delta_\alpha = P_\alpha\{Y = Y'\}/2 = \frac{p}{4-2p}.$$

Remember that $\Delta_{\alpha(x_i,y_i)} = 1 - F(x_i) - G(y_i) + H(x_i, y_i) + H(x_i, y_i - 1)$, which is maximised by M and minimised by W. We compute numerically this term as well as the other ones.

Since $p = 0.01140351$, it follows that

$$\sup \Delta_\alpha = 0.9971327743 \text{ and } \inf \Delta_\alpha = 0.002867225719.$$

Besides, lower and upper bounds on $\Delta_{(\alpha,\hat{\alpha})}$ are given, respectively, by .53522497 and .85383703, whereas $\Delta_{\hat{\alpha}} = .82084156$.

We get Table 6.2 about upper and lower bounds, for different choices of β.

From Table 6.2, we can see that upper and lower bounds are getting closer and closer if β diminishes. Besides, the most relevant finding is that, even for

Table 6.1: Sannio Matese earthquakes data

X	Y	X	Y	X	Y	X	Y
11.58680	90	1.40109	143	0.99848	15	0.48031	1
1.23643	35	3.52873	88	0.51621	11	1.88481	39
10.03350	179	7.97373	192	1.75758	128	0.80690	58
0.00100	0	0.04216	10	1.52169	120	0.18979	11
0.36307	23	4.16688	236	0.82927	44	1.74428	44
0.75327	37	1.27373	62	7.96161	461	2.65489	94
0.08134	2	2.04751	77	2.78165	51	0.28868	22
2.88668	118	7.90076	140	11.35160	69	2.22582	16
7.42631	102	0.00002	0	15.35818	533	0.99172	16
1.90574	92	0.00532	21	0.00016	1		

Since the marginal distribution of X is exponential with parameter $\lambda^* p^* = 0.3278477195$, then $F(x) = 1 - e^{-0.3278477195x}$, for all $x \in \Re^+$. The marginal distribution of Y is geometric with parameter $p^* = 0.01140351$, so that $G(y) = 1 - 0.98859649^{y+1}$, for all $y \in \mathcal{N}$.

Different values of β will be considered in the next sections.

6.2 Statistical analysis

6.2.1 Fréchet class

Consider the family of Dirichlet processes with parameter $\eta = \beta\alpha, \beta > 0$ and $\alpha \in \Gamma(F, G)$, where F and G are given as in Section 6.1. Therefore the prior Fréchet bounds are given by

$$M(x,y) = \begin{cases} F(x) & x \leq 0.03498279(y+1) \\ G(y) & x > 0.03498279(y+1) \end{cases},$$

and by

as chosen by Polya trees (see Lavine [47]).

Here we suppose that the distribution of (X, Y) is chosen according to a Dirichlet process and that it is possible to express an opinion on the marginal distributions of X and Y, but not about their dependence. Therefore, we assume that the marginals of X and Y are like the ones they would have under the Poisson model, i.e. exponentially and geometrically distributed, respectively. In addition, the parameters λ and p are chosen as the maximum likelihood estimators for the Poisson model.

The likelihood function for n observations $(x_1, y_1), \ldots, (x_n, y_n)$ is given by

$$l(\lambda, p) = \prod_{i=1}^{n} f(x_i, y_i) = \lambda^n p^n e^{-\lambda \sum x_i} \lambda^{\sum y_i} (1-p)^{\sum y_i} \prod x_i^{y_i} / \prod y_i!,$$

so the loglikelihood is given, apart from an additive constant, by

$$L(\lambda, p) = (n + \sum y_i) \log \lambda - \lambda \sum x_i + n \log p + \sum y_i \log(1-p).$$

By considering the derivatives of the loglikelihood with respect to the variables λ and p, it follows that the maximum likelihood estimators are given by

$$\lambda^* = \frac{n + \sum y_i}{\sum x_i} \text{ and } p^* = \frac{n}{\sum y_i + n},$$

i.e. by the ratio between the total number of earthquakes and the elapsed time between the first and the last major ones and the proportion of major earthquakes among all those occurred, respectively.

By looking at Table 6.1, it follows that $n = 39, \sum x_i = 118.9577$ and $\sum y_i = 3381$, so that

$$\lambda^* = 28.74972 \text{ and } p^* = 0.01140351.$$

The preceding model is very similar to the one considered by Ruggeri [57], in which the magnitude of major earthquakes was also considered. In that paper, a Bayesian parametric approach was followed and predictive distributions of interoccurrence times of major earthquakes, of magnitudes and number of minor earthquakes between two major ones were computed, along with the posterior distributions of the parameters such quantities depend on.

Here we relax the hypothesis that the earthquakes occur according to a Poisson process. In fact, the Poisson model should be discarded because of the analysis performed in the paper by Ladelli, Mitrione, Ruffoni and Ruggeri [46] where several point process models, including the simple Poisson process, were considered to describe the same earthquake occurrences, and the parameters were estimated by means of likelihood methods, while the Akaike Information Criterion and spectral analysis were used to compare models. An epidemic-type model was chosen as the best among all those under consideration, as confirmed by a residual analysis.

Restricting the research of a model to the parametric ones could be practical and sometimes convincing, but we think that a nonparametric approach could sometimes offer new insights in analysing the data, specially if combined with the prior knowledge, like in Bayesian nonparametrics. Here we are analysing a large area with a great number of occurrence; experts in seismology convene that such features are likely to recommend a Poisson process to describe the occurrence of earthquakes. Since we do not want to commit ourselves to such an hypothesis, we present an approach which is related to the Poisson model, but allows more flexibility and data dependence on the distribution function and the dependence between the random variable. Of course, the model could be strongly improved by allowing for absolutely continuous probability measures, consider just one earthquake, the strongest, as the main shock in any sequence lasting one week. We consider the earthquakes occurring since 1860, as suggested by an exploratory data analysis, performed by Ladelli, Mitrione, Ruffoni and Ruggeri [46], which identified three different behaviours of the occurrence time process since 1120. Also, we consider, as the first interoccurrence time, the elapsed time between the first and second earthquake.

As discussed before, the assumption of independence between subsequent earthquakes is arbitrary, but, nonetheless, quite well accepted in many papers, as well as the assumption that two or more earthquakes can occur at the same time so that it is very popular to describe the earthquakes occurrences by means of a Poisson process (see, e.g., Campbell [14]). Therefore, consider a Poisson process, with constant rate λ (see, e.g., Çinlar [19] on Poisson processes). Suppose that each earthquake has probability p of being a major one (and $1-p$ of being a minor one), so that we can decompose the Poisson process into two independent ones with rate λp and $\lambda(1-p)$, corresponding, respectively, to major and minor earthquakes. It follows that the interoccurrence times X are exponentially distributed with mean $1/\lambda p$, while, conditionally on the time x (realisation of X), Y is Poisson distributed, with mean $\lambda(1-p)x$. It is straightforward to prove that Y is marginally geometrically distributed, with parameter p, because, for $y \in \mathcal{N}$,

$$f(y) = \int_0^\infty f(y|x)f(x)dx = \int_0^\infty e^{-\lambda(1-p)x}\frac{[\lambda(1-p)x]^y}{y!}\cdot \lambda p e^{-\lambda p x}dx = p(1-p)^y,$$

where f denotes the density function of a random variable.

The joint density of (X, Y) is given by

$$f(x,y) = f(y|x)f(x) = \lambda p e^{-\lambda x}\frac{[\lambda(1-p)x]^y}{y!}.$$

statistical estimation of the "seismic hazard function", i.e. the hazard function of the renewal process of the occurrence times. More recent references are the papers by Ladelli, Mitrione, Ruffoni and Ruggeri [46], in which earthquake occurrences are modelled by means of point processes, and by Ruggeri [57], who proposed a model to analyse the interoccurrence times of major earthquakes, their magnitudes and the number of minor earthquakes between two major ones from a Bayesian point of view, making comparison between three geographically homogeneous areas in Sannio Matese.

In this chapter we will analyse the earthquakes in Sannio Matese, an area in southern Italy subject to a consistent and sometimes very disruptive seismic activity, as demonstrated in Casamicciola (1883), Avezzano (1915) and Irpinia-Lucania (1980).

6.1 Description of the model

In this paper, we consider the random variables X and Y which denote, respectively, the interoccurrence times (in years) of major earthquakes (i.e. with magnitude not smaller than 5), and the number of minor earthquakes occurring in a given area since the previous major one. The magnitude M used to be computed from the intensity I by using a relation due to Karnik [44], given, in Sannio Matese, by $M = 0.51I + 1$, whereas it has been instrumentally recorded in more recent years. We consider our data as a sample from $Z = (X, Y)$ and assume that $Z_1, Z_2 \ldots$ are independent. Although accepted in previous works, such an assumption is unrealistic, because sometimes earthquakes are accompanied by sequences of foreshocks and aftershocks, all of minor intensity. Since the presence of foreshocks and aftershocks can be hard to recognise, we

Chapter 6

Earthquakes in Sannio Matese

Earthquakes are a problem in Italy, because they affect many areas of the country (Friuli, Garfagnana, Sannio Matese, Arco Calabro, etc.), causing many deaths and disruptions. For this reason, many Italian institutions, including Consiglio Nazionale delle Ricerche (CNR) and universities, have been studying the phenomenon. One major project has been the compilation of a catalogue, prepared by the *Gruppo Catalogo dei Terremoti dell'Istituto per la Geofisica della Litosfera del CNR*, containing all the recorded earthquakes since 1120, using current and historical data (e.g. church records). For each earthquake, the catalogue contains occurrence time (often accurate within seconds), latitude and longitude, intensity (i.e. visible effects), magnitude (sometimes recorded instrumentally, otherwise estimated from the intensity), the position in the maps of the *Istituto Geografico Militare* and the name of the place in which it occurred.

The data collected in the catalogue have been the object of some statistical analysis. See, for example, Betrò, Garavaglia, Guagenti, Rotondi and Tagliani [10], where a survey of the results is presented, along with the study of the

A, that $\dfrac{\eta_2(A)}{\eta_1(A)} e^{(\eta_1(A)-\eta_2(A))/\eta_1(A)} < 1$.

To complete the proof, observe that the function Y achieves its supremum a posteriori for a subset A^* such that $\eta_0^*(A^*) = x$ and $\eta^*(A^*) = \varphi(x)$. Since the prior Y is strictly greater than the posterior one, its prior supremum is greater than the prior Y evaluated at A^*, so it is greater than the posterior supremum. It follows that the distance is strictly decreasing whence a sample is given. □

Theorem 5.7.5 $\lim_{n\to\infty} d_{KL}(P, Q) = 0$ *a.s.*.

PROOF. Like in the proof of Theorem 5.7.3, we see that

$$Y(\eta_1(A) + \sum_{i=1}^{n} \delta_{Z_i}(A), \eta_2(A) + \sum_{i=1}^{n} \delta_{Z_i}(A)) = o(1) \text{ a.s.},$$

applying both the same asymptotic formula and the following one:

$$\Psi(z) \sim \log z - \frac{1}{2z} - \sum_{n=1}^{\infty} \frac{B_{2n}}{2nz^{2n}}, z \to \infty,$$

where B_n are the Bernoulli numbers (see Abramowitz and Stegun [1], p.259).

□

PROOF. Given a sample Z_i and a subset $A \in \mathcal{A}$, then $\delta_{Z_i}(A)$ can be seen as a Bernoulli random variable having mean $\tilde{\eta}(A)$, the "true" probability of A (to avoid triviality, we suppose that $0 < \tilde{\eta}(A) < 1$). Because of the Strong Law of Large Numbers, it follows that $\sum_{i=1}^{n} \delta_{Z_i}(A)/n \to \tilde{\eta}(A)$ a.s., so that $\sum_{i=1}^{n} \delta_{Z_i}(A)$ is unbounded (and similarly $n - \sum_{i=1}^{n} \delta_{Z_i}(A)$).

Applying the asymptotic formula, presented in Abramowitz and Stegun [1], p. 257,

$$\Gamma(az+b) \sim \sqrt{2\pi} e^{-az} (az)^{az+b-1/2}, z \to \infty,$$

it follows that $Y(\eta_1(A) + \sum_{i=1}^{n} \delta_{Z_i}(A), \eta_2(A) + \sum_{i=1}^{n} \delta_{Z_i}(A)) \sim 1$ a.s., completing the proof. □

Theorem 5.7.4 *$d_{KL}(P^*, Q^*) < d_{KL}(P, Q)$ where P^* and Q^* are the Dirichlet processes obtained by updating the processes P and Q after observing a sample of size n.*

PROOF. We consider a sample of size 1 because the case n is obtained by reiteratively applying the proof for the case 1. We do not consider the case $\eta_1(A) = \eta_2(A)$ which gives a null distance, both a priori and a posteriori. We prove, first of all, that

$$e^{Y(\eta_1(A)+\delta_{Z_1}(A), \eta_2(A)+\delta_{Z_1}(A))} < e^{Y(\eta_1(A), \eta_2(A))},$$

where $Y(\eta_1(A), \eta_2(A)) = \int_0^1 f_{\eta_1(A)}(y) \log \frac{f_{\eta_1(A)}(y)}{f_{\eta_2(A)}(y)} dy$. Such inequality is always satisfied because simple computations show that proving it is equivalent to prove, when $Z_1 \notin A$, that $\frac{\beta - \eta_2(A)}{\beta - \eta_1(A)} e^{(\eta_2(A) - \eta_1(A))/(\beta - \eta_1(A))} < 1$ and, when $Z_1 \in$

PROOF. We consider a sample of size 1 because the case n is obtained by reiteratively applying the proof for the case 1. We do not consider the case $\eta_1(A) = \eta_2(A)$ which gives a null distance, both a priori and a posteriori. We prove, first of all, that

$$Y(\eta_1(A) + \delta_{Z_1}(A), \eta_2(A) + \delta_{Z_1}(A)) > Y(\eta_1(A), \eta_2(A)),$$

where $Y(\eta_1(A), \eta_2(A)) = \int_0^1 \sqrt{f_{\eta_1(A)}(y) f_{\eta_2(A)}(y)}dy$. Note, in fact, that

$$\frac{Y(\eta_1(A) + \delta_{Z_1}(A), \eta_2(A) + \delta_{Z_1}(A))}{Y(\eta_1(A), \eta_2(A))} =$$

$$= \frac{\Gamma((\eta_1(A) + \eta_2(A))/2 + \delta_{Z_1}(A))}{\Gamma((\eta_1(A) + \eta_2(A))/2)} \frac{\Gamma(\beta + 1 - (\eta_1(A) + \eta_2(A))/2 - \delta_{Z_1}(A))}{\Gamma(\beta - (\eta_1(A) + \eta_2(A))/2)} \cdot$$

$$\cdot \sqrt{\frac{\Gamma(\eta_1(A))}{\Gamma(\eta_1(A) + \delta_{Z_1}(A))} \frac{\Gamma(\beta - \eta_1(A))}{\Gamma(\beta + 1 - \eta_1(A) - \delta_{Z_1}(A))} \frac{\Gamma(\eta_2(A))}{\Gamma(\eta_2(A) + \delta_{Z_1}(A))}} \cdot$$

$$\cdot \sqrt{\frac{\Gamma(\beta - \eta_2(A))}{\Gamma(\beta + 1 - \eta_2(A) - \delta_{Z_1}(A))}}$$

The above quantity equals $\dfrac{\beta - (\eta_1(A) + \eta_2(A))/2}{\sqrt{(\beta - \eta_1(A))(\beta - \eta_2(A))}}$ when $Z_1 \notin A$ and $\dfrac{(\eta_1(A) + \eta_2(A))/2}{\sqrt{\eta_1(A)\eta_2(A))}}$ when $Z_1 \in A$. Both quantities are greater than one if and only if $(\eta_1(A) - \eta_2(A))^2 > 0$, which is always true.

To complete the proof, observe that the function Y achieves its infimum a posteriori for a subset A^* such that $\eta_0^*(A^*) = x$ and $\eta^*(A^*) = \varphi(x)$. Since the prior Y is strictly less than the posterior one, its prior infimum is lower than the prior Y evaluated at A^*, so it is lower than the posterior infimum. It follows that the distance is strictly decreasing whence a sample is given. □

Theorem 5.7.3 $\lim_{n \to \infty} d_H(P, Q) = 0$ *a.s.*.

after observing a sample of size n. Then it follows that

$$\rho(\eta^*, g) = \rho(\eta, g) + ng(1).$$

PROOF. Using the notations described in Appendix B, we know that

$$H(y) = \eta_0\left(\{\theta \in \Theta : h(\theta) \leq y\}\right) \text{ and } c_x \;=\; \inf\{y \in \Re : H(y) \geq x\},$$

whereas

$$H^*(y) = \eta_0^*\left(\{\theta \in \Theta : h(\theta) \leq y\}\right) = \begin{cases} H(y) & 0 \leq x < 1 \\ H(y) + n & x \geq 1 \end{cases}$$

and

$$c_x^*(x) = \inf\{y \in \Re : H^*(y) \geq x\} = \begin{cases} c(x) & x < H(1) \\ 1 & H(1) \leq x \leq H(1) + n \\ c(x-n) & x > H(1) + n \end{cases}.$$

The result is proved by observing that

$$\eta_s(\Theta) = \eta_s^*(\Theta) \text{ and } \int_{[0,\infty)} g(t)dH^*(t) = \int_{[0,\infty)} g(t)dH(t) + ng(1).$$

□

Consider now the distance between Dirichlet processes as described in Section 5.3; as expected, any sample decreases such a distance. In the next, we suppose that data come from the "true" distribution $\tilde{\eta}$ and that the support of $\tilde{\eta}$ contains the support of the parameters of the Dirichlet processes.

Theorem 5.7.2 $d_H(P^*, Q^*) < d_H(P, Q)$ *where* P^* *and* Q^* *are the Dirichlet processes obtained by updating the processes* P *and* Q *after observing a sample of size* n*.*

about $Z = (X, Y)$, the Bayes estimators of variance and means of X and Y are constant (see Section 4.1.2). We are thus facing the nice situation of having a large class, $\Gamma(F, G)$, in which robustness is achieved. Upper and lower bounds on the covariance have been found, too, in Section 4.1.2.

Following a suggestion in Ruggeri [61], it could be checked if the Bayes estimator of the distribution, $\eta(-\infty, x)/\eta(\Re)$, is within a prespecified band, maybe around a baseline distribution function. In the Fréchet class, bounds are given by $\hat{W}$ and $\hat{M}$.

5.7 Comparing prior and posterior distances

So far, we have measured the distance between quantities of interest, presenting results which are valid both before and after observing a sample; now we want to check if the sample influences, and in which direction, the above distances.

First of all, it is worth considering the coefficients of divergence

$$\rho(\eta, g) = \int_{[0,\infty)} g(t) dH_\eta(t) + \eta_s(\Theta) \lim_{t \to \infty} \{g(t)/t\},$$

described in Appendix B, when applied to compare the distance between a parameter $\eta \in \Gamma$ and a baseline η_0. It follows that prior and posterior distances coincide for most of the well-known indices (Kullback-Leibler, χ^2–divergence and Pietra) but not for Gini's one, that is for all indices for which $g(1) = 0$, i.e. which compare probability measures giving no weight where the two measures coincide.

Theorem 5.7.1 *Let $\rho(\eta, g)$ be the coefficient of divergence of η w.r.t. η_0 and let $\rho(\eta^*, g)$ be the coefficient of divergence of η^* w.r.t. η_0^*, updated parameters* the Dirichlet process in a class Γ, so that

$$\mathcal{P}(\Pi) = \frac{\int_0^x y^{\eta(A)-1}(1-y)^{\beta-\eta(A)-1}dy}{\int_0^1 y^{\eta(A)-1}(1-y)^{\beta-\eta(A)-1}dy}.$$

Applying Lemma 5.4.1, it follows that upper and lower bounds on $\mathcal{P}(\Pi)$ are achieved for $\eta(A)$ equal to $\inf_{\eta\in\Gamma}\eta(A)$ and $\sup_{\eta\in\Gamma}\eta(A)$, respectively. Another interesting case is given by $A=(\infty, y)$, so that we get a class of random distribution functions $\{F : F(y) < x\}$, e.g., for $x = 1/2$, we have the class of all the distributions whose median is greater than y.

Bounds can be easily computed for the Fréchet class described in Example 5.4.2, too.

5.6 Distance between Bayes estimators

In the literature on Bayesian nonparametrics, many Bayes estimators of quantities of interest have been presented. It could be interesting to investigate how much they change as the parameter η of a Dirichlet process varies in a class Γ. This problem can be reduced to a usual one in parametric robustness. Consider the estimation of the mean and the estimation of a distribution function, both under squared loss, solved in Ferguson [30]. We consider the no-sample problem, being similar to the one with data. The Bayes estimator of the mean is given by $\int_{\Re} x d\alpha(x)$ and η could be in most of the classes considered in the parametric literature, for which methods for computing bounds are well known. As an example, let the density η be bounded within the densities l and u, then the upper bound on the Bayes estimator is given by $\hat{\eta} \equiv l$ on $(-\infty, x]$ and $\hat{\eta} \equiv u$ on (x, ∞), x being determined by $\hat{\eta}(\Re) = \eta(\Re)$. In the Fréchet class

Subsets of probability measures can be defined by means of the random functionals considered by Cifarelli and Regazzini [17], [18]. Provided that Condition C.1.2 is satisfied, then it possible to consider the random functional

$$Y_\psi = \int_\Re \psi(x) P(dx),$$

whose distribution $\mathcal{M}$ follows from Theorem C.1.1 and Corollary C.1.1. We can now consider the class $\Gamma = \{\Pi \in \mathcal{P} : Y_\psi \in B\}$.

Theorem 5.5.3

$$\mathcal{P}(\Gamma) = \int_B d\mathcal{M}(x).$$

Consider now a family $\mathcal{R}$ of Dirichlet processes, where the parameter η belongs to a class Δ. Like in the parametric Bayesian robust analysis, the sensitivity to the changes in η is measured by considering upper and lower bounds on the probabilities of subsets of the space of all probabilities. In particular, it is possible to consider α in the Fréchet class $\Gamma(F, G)$, observing that Theorem 5.5.3 does not apply to such case because it holds only for measures on $\Re$. Because of the nature of the Dirichlet process, it is possible to define some classes of probability measures $\Pi = \{P : P(A) \in B\}$, where $A \in \mathcal{A}$ and B is a Lebesgue measurable subset in $[0, 1]$. Similar classes could be defined by asking either that the random probabilities of a finer partition of $\mathcal{X}$ belong to some subset or the random functional, considered by Cifarelli and Regazzini [17], [18], takes value on some subset. Being $P(A)$ Beta distributed, it is easy to compute the probability of Π. As in parametric robustness, it is worthwhile to compute upper and lower bounds on the probability $\mathcal{P}(\Pi)$.

As an example, we could take $\Pi = \{P : P(A) \leq x\}$ and the parameter η of

5.5 Distance between sets of probability measures

In this section, we compare the probabilities given by a class of Dirichlet processes to some subspaces of the space $\mathcal{P}$ of all the probability measures defined on the space $(\mathcal{X}, \mathcal{A})$. Given the subsets $A \in \mathcal{A}$ and $B \in \mathcal{B}$, where $\mathcal{B}$ is the Borel σ–field on $[0,1]$, consider the subspace $\Gamma = \{\Pi \in \mathcal{P} : \Pi(A) \in B\}$. Let P be a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter η, then the random probability $P(A)$ is Beta distributed with parameters $(\eta(A), \beta - \eta(A))$, with density $f_{\eta(A)}$. The following result can therefore be easily proved.

Theorem 5.5.1

$$\mathcal{P}(\Gamma) = \mathcal{P}(P(A) \in B) = \int_B f_{\eta(A)}(x)dx.$$

More generally, it is possible to consider a measurable partition $A_1, \ldots, A_k$ of $\mathcal{X}$, and the measurable subsets $B_1, \ldots, B_k$ in $\mathcal{B}$, and define the class

$$\Gamma = \{\Pi \in \mathcal{P} : \Pi(A_i) \in B_i, i = 1, \ldots, k\}.$$

From Definition 2.2.1, it follows that the joint distribution of the random probabilities $(P(A_1), \ldots, P(A_k))$ is Dirichlet with parameters $(\eta(A_1), \ldots, \eta(A_k))$, with density $f_{\eta(A_1),\ldots,\eta(A_k)}$. The following result can therefore be easily proved.

Theorem 5.5.2

$$\mathcal{P}(\Gamma) = \int_{B_1 \times \ldots \times B_k} f_{\eta(A_1),\ldots,\eta(A_k)}(x_1, \ldots, x_k)dx_1 \ldots dx_k.$$

It should be observed that Theorem 5.5.1 is a special case of Theorem 5.5.2, when $k = 2$, but it has been presented separately to have a handy reference in the forthcoming examples.

Proof. Consider any $\eta \in \Gamma$ such that $\eta(A) > \eta_0(A)$. Let Q be the corresponding Dirichlet process, then it follows from Lemma 5.4.2 that the c.f. of $Q(A)$ w.r.t $P_0(A)$ is given by $x = \int_0^y f_{\eta_0}(t)dt$ and $\varphi(x) = \int_0^y f_\eta(t)dt$.

From Lemma 5.4.1, it follows that $\int_0^y f_\eta(t)dt$ is decreasing in η, so that it is minimised by $\eta_2(A)$. A similar argument can be applied to η's such that $\eta(A) < \eta_0(A)$. □

Example 5.4.1 Take the ε–contaminated class $\Gamma = \{\eta = (1-\varepsilon)\eta_0 + \varepsilon Q, Q \in \mathcal{M}_\beta\}$, where $\mathcal{M}_\beta$ is the class of all finite measures Q such that $Q(\mathcal{X}) = \eta(\mathcal{X}) = \beta$. In this case Q_1 and Q_2 are such that $\eta_1(A) = 0$ and $\eta_2(A) = \beta$, respectively. It follows from Proposition 2.2.1 that $Q_1(A) = 0$ a.s. and $Q_2(A) = \beta$ a.s., i.e. $Q_1(A)$ and $Q_2(A)$ are Dirac measures concentrated at 0 and β, respectively. Therefore their c.f.'s w.r.t. $P_0(A)$ are such that $\varphi_1(x) = \varphi_2(x) = \hat{\varphi}(x) = 0$, for all $x \in [0,1]$.

The same $\hat{\varphi}$ is obtained if we consider the parameters η and η_0 updated after observing a sample Z_1 of size 1 (extension to larger sample is trivial). If $Z_1 \in A$, then the posterior $Q_2^*(A)$ is a Dirac measure concentrated at $\beta + 1$; otherwise, the posterior $Q_1^*(A)$ is a Dirac measure concentrated at 0. In both cases the corresponding c.f. w.r.t. the updated η_0^* is equal to zero everywhere.

Example 5.4.2 Consider all the parameters $\eta = \beta\gamma$, $\beta > 0$, with the probability measure γ in the Fréchet class $\Gamma(F, G)$ and, as a baseline parameter, any $\eta_0 = \beta\gamma_0$ with γ_0 in Γ, e.g. the independent one with joint distribution $F(x)G(y)$. Take the subset $A = (-\infty, x] \times (-\infty, y]$, then it follows that Q_1 and Q_2 are such that $\eta_1(A) = \beta W(x,y)$ and $\eta_2(A) = \beta M(x,y)$. Update η and η_0 after observing a sample of size n, then the updated Q_1^* and Q_2^* are such that $\eta_1^*(A) = (\beta + n)\hat{W}(x,y)$ and $\eta_2^*(A) = (\beta + n)\hat{M}(x,y)$.

Lemma 5.4.2 *Let $0 < \eta_1, \eta_2 < \beta$, and let P and Q be Beta distributions with parameters $(\eta_1, \beta - \eta_1)$ and $(\eta_2, \beta - \eta_2)$, respectively, and densities f_{η_1} and f_{η_2}; then the c.f. of Q w.r.t. P can be computed, for any $y \in [0,1]$, as*

$$\begin{aligned} x &= \int_0^y f_{\eta_1}(t)dt, \quad \varphi(x) = \int_0^y f_{\eta_2}(t)dt \quad \eta_2 \geq \eta_1 \\ x &= \int_y^1 f_{\eta_1}(t)dt, \quad \varphi(x) = \int_y^1 f_{\eta_2}(t)dt \quad \eta_2 \leq \eta_1. \end{aligned}$$

Proof. It can be easily shown that the likelihood ratio is given by

$$h(\theta) = \frac{f_{\eta_2}(\theta)}{f_{\eta_1}(\theta)} = K \left(\frac{\theta}{1-\theta} \right)^{\eta_2 - \eta_1},$$

where K is a constant independent of θ. As $h(\theta)$ is increasing (decreasing) for $\eta_2 > \eta_1$ $(\eta_2 < \eta_1)$, it follows, using the same notation as in Appendix B, that L_x has the form $[0, y_x]$ $([y_x, 1])$, for all $x \in [0,1]$. $\square$

We can now consider the distribution function of the probability $P(A)$ of $A \in \mathcal{A}$, when P is chosen from a Dirichlet process. Suppose there exists a baseline process P_0 with parameter η_0 and a class of processes Q with parameters η in Γ, with $\eta_0 \in \Gamma$. We compare the distributions of $P(A)$ under the processes Q with the one under P_0 by using their c.f.'s, as presented in Fortini and Ruggeri [37]. For each $x \in [0,1]$, we look for the lowest c.f. $\hat{\varphi}(x)$, i.e. for the minimum probability, under the distributions of $Q(A)$, of all subsets having probability x under the distribution of $P_0(A)$.

Theorem 5.4.1 *Let Q_1 and Q_2 be the Dirichlet processes with parameters η_1 and η_2, respectively, such that $\eta_1(A) = \inf_{\eta \in \Gamma} \eta(A)$ and $\eta_2(A) = \sup_{\eta \in \Gamma} \eta(A)$. Let φ_i be the c.f. of $Q_i(A)$ w.r.t. $P_0(A)$, $i = 1, 2$, then, for all $x \in [0,1]$, $\hat{\varphi}(x) = \min\{\varphi_1(x), \varphi_2(x)\}$.*

5.4 Distance between distributions of probabilities

Whereas the parametric robust Bayesian approach is often interested in finding upper and lower bounds on posterior set probabilities and expectations in the parameter space, the nonparametric approach must deal with the changes in the distribution of probabilities and expectations in the sample space. In this paper we focus only on set probabilities, simply mentioning that ranges of distributions could be found for all random functionals for which Cifarelli and Regazzini [17], [18] have computed the distribution functions (see Appendix C for a short summary of their results).

Lemma 5.4.1 *Let $0 < \eta < \beta$, then the function*

$$Y(\eta) = \frac{\int_0^x y^{\eta-1}(1-y)^{\beta-\eta-1}dy}{\int_0^1 y^{\eta-1}(1-y)^{\beta-\eta-1}dy}$$

is strictly decreasing in η.

PROOF. Let $L(y) = \log \dfrac{y}{1-y}$ and $f_\eta(y) = y^{\eta-1}(1-y)^{\beta-\eta-1}$, then

$$Y'(\eta) = \frac{Z_\eta(x)}{\left(\int_0^1 f_\eta(y)dy\right)^2},$$

with $Z_\eta(x) = \int_0^x L(y)f_\eta(y)dy \int_x^1 f_\eta(y)dy - \int_x^1 L(y)f_\eta(y)dy \int_0^x f_\eta(y)dy.$

For any $\eta \in (0,\beta)$, it follows that $\dfrac{\partial Z_\eta(x)}{\partial x} = 0$ inside $(0,1)$ at the unique point $\hat{x}_\eta$ such that $L(\hat{x}_\eta) = \mathcal{E}_\eta L$, being $\hat{x}_\eta$ a minimum, since $Z_\eta(1/2) < Z_\eta(0) = Z_\eta(1) = 0$. It follows that, for all η's, Z_η is negative, except for $x = 0, 1$, so that $Y(\eta)$ is strictly decreasing in η. □

Table 5.2: Kullback-Leibler divergence

γ	.0001	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.
Distance	8.76	1.9	1.2	.79	.53	.35	.21	.12	.05	.01	0.

Distances are computed for different values of γ and shown in Table 5.2.

Example 5.3.2 (continued) It can be shown that $d_{KL}(P,Q)$ is unbounded also in this case.

As for Hellinger distance, we can prove a Theorem, similar to Theorem 5.3.2, which simplify the computation of the distance when considering the class Λ of all c.f.'s of the parameters $\eta \in \Gamma$ w.r.t. η_0.

Theorem 5.3.5 *For any $x \in [0,1]$, define $\hat{\varphi}(x) = \inf_{\varphi \in \Lambda} \varphi(x)$. Then it follows that*

$$\sup_{Q \in \Pi} d_{KL}(P_0, Q) = \sup_{0 \le x \le \beta} Y(x, \hat{\varphi}(x)).$$

As before about the Hellinger distance, the class Γ could be described by means of the c.f..

Example 5.3.3 (continued) For the ε–contaminated class with $\beta = 1$, we obtain that $\lim_{x \to \beta} Y(x, (1-\varepsilon)x) = \infty$ and therefore there exists Q such that $d_H(P_0, Q) = \infty$ (the proof is the same as in Example 5.3.1).

Example 5.3.4 (continued) In the total variation neighbourhood of η_0, it can be proved that, given $0 < x < \varepsilon$, then $\lim_{t \to 0} Y(x,t) = +\infty$ so that there exists Q such that $d_{KL}(P_0, Q) = \infty$.

Since it can be proved that $Y(x, \beta - \varphi(\beta - x)) = Y(\beta - x, \varphi(\beta - x))$, it follows that it is sufficient to consider $Y(x, \varphi(x))$. $\square$.

Therefore the distance between two processes is found by computing the c.f. $\varphi(x)$ and then maximising the function $Y(x, \varphi(x))$, as in the following examples.

Example 5.3.1 (continued) For $x \leq .5$ and $\beta = 1$, it follows that

$$Z_x(\gamma) = \frac{\partial Y(x, \varphi(x))}{\partial x} = -\gamma\pi \cot \gamma\pi x + \gamma\pi \cot \pi x + \frac{(1-\gamma)\pi^2 x}{\sin^2 \pi x},$$

so that $Z_x(1) = 0$ for all $x \in [0, 1]$. Besides

$$\frac{\partial Z_x(\gamma)}{\partial \gamma} = -\pi \cot \gamma\pi x + \frac{\gamma\pi^2 x}{\sin^2 \gamma\pi x} + \pi \cot \pi x - \frac{\pi^2 x}{\sin^2 \pi x},$$

which equals 0 at $\gamma = 1$. The quantity within brackets in

$$\frac{\partial^2 Z_x(\gamma)}{\partial \gamma^2} = \frac{2\pi^2 x}{\sin^3 \gamma\pi x}\{\sin \gamma\pi x - \gamma\pi x \cos \gamma\pi x\}$$

is always positive (obvious for $\gamma x > 1/2$, whereas otherwise we should look at $t < \tan t$ for $0 \leq t < \pi/2$).

Therefore $\frac{\partial^2 Z_x(\gamma)}{\partial \gamma^2} > 0$ implies that $\frac{\partial Z_x(\gamma)}{\partial \gamma}$ is increasing and, because of $\left\{\frac{\partial Z_x(\gamma)}{\partial \gamma}\right\}_{\gamma=1} = 0$, that it negative for $\gamma < 1$. As a consequence, then Z_x is decreasing as a function of γ and, because of $Z_x(1) = 0$, it is positive for any $\gamma \in (0, 1)$, given any $x \in [0, 1]$. Therefore $\frac{\partial Y(x, \varphi(x))}{\partial x}$ is positive for $x \leq .5$ while it can be similarly proved that it is negative for $x > .5$, so that $Y(x, \varphi(x))$ achieves its maximum value at $x = .5$. Finally, it can be shown that $\lim_{\gamma \to 0} Y(x, \varphi(x)) = \infty$ so that $d_{KL}(P, Q) = \infty$.

$$= \int_0^1 \frac{y^{\eta_1(A)-1}(1-y)^{\beta-\eta_1(A)-1}\Gamma(\beta)}{\Gamma(\eta_1(A))\Gamma(\beta-\eta_1(A))} \cdot$$

$$\cdot \log \frac{y^{\eta_1(A)-1}(1-y)^{\beta-\eta_1(A)-1}\Gamma(\beta)\Gamma(\eta_2(A))\Gamma(\beta-\eta_2(A))}{y^{\eta_2(A)-1}(1-y)^{\beta-\eta_2(A)-1}\Gamma(\eta_1(A))\Gamma(\beta-\eta_1(A))\Gamma(\beta)} dy$$

$$= -\log\Gamma(\eta_1(A)) - \log\Gamma(\beta-\eta_1(A)) + \log\Gamma(\eta_2(A)) + \log\Gamma(\beta-\eta_2(A)) +$$

$$+(\eta_1(A)-\eta_2(A))\{\Psi(\eta_1(A)) - \Psi(\beta-\eta_1(A))\}.$$

We now compute the distance between any Dirichlet process Q with parameter η_2 from a baseline Dirichlet process P with parameter η_1.

Theorem 5.3.4 $d_{KL}(P,Q) = \sup_{0 \le x \le \beta} Y(x, \varphi(x))$ *where* φ *is the c.f. of* η_2 *w.r.t.* η_1.

Proof. Consider $\mathcal{A}_x = \{A \in \mathcal{A} : \eta_1(A) = x\}$, $0 \le x \le \beta$; we want to find $A \in \mathcal{A}_x$ which maximises $Y(x, \eta_2(A))$. It follows that

$$\frac{\partial Y}{\partial \eta_2(A)} = \frac{\Gamma'(\eta_2(A))}{\Gamma(\eta_2(A))} - \frac{\Gamma'(\beta-\eta_2(A))}{\Gamma(\beta-\eta_2(A))} - \mathcal{E}_{\eta_1(A)} L,$$

where $\mathcal{E}_\eta L$ denotes the expected value of $L = \log \frac{T}{1-T}$, when T is Beta distributed with parameters $(\eta, \beta-\eta)$.

Furthermore,

$$\begin{aligned}
\frac{\partial^2 Y}{\partial \eta^2(A)} &= \frac{\Gamma''(\eta_2(A))\Gamma(\eta_2(A)) - \left(\Gamma'(\eta_2(A))\right)^2}{\left(\Gamma(\eta_2(A))\right)^2} - \\
&\quad - \frac{-\Gamma''(\beta-\eta_2(A))\Gamma(\beta-\eta_2(A)) + \left(\Gamma'(\beta-\eta_2(A))\right)^2}{\left(\Gamma(\eta_2(A))\right)^2} = \\
&= Var_{\eta_2(A)} \log T + Var_{\beta-\eta_2(A)} \log T > 0.
\end{aligned}$$

The convexity of Y implies that Y takes its maximum value at either $\inf_{A \in \mathcal{A}_x} \eta_2(A)$ or $\sup_{A \in \mathcal{A}_x} \eta_2(A)$, i.e., as discussed in Appendix B, at either $\varphi(x)$ or $\beta - \varphi(\beta - x)$. The same argument can be repeated for any $x \in [0, \beta]$.

Theorem 5.3.3 $\sup_{\{B_1,\ldots,B_k\}} d(\{P(B_1),\ldots,P(B_k)\},\{Q(B_1),\ldots,Q(B_k)\})$ *is a nondecreasing function of* k*.*

PROOF. It follows from

$$\frac{\int_0^1 \sqrt{f_{\eta_{10},\eta_{11},\ldots,\eta_k} f_{\gamma_{10},\gamma_{11},\ldots,\gamma_k} dy}}{\int_0^1 \sqrt{f_{\eta_1,\ldots,\eta_k} f_{\gamma_1,\ldots,\gamma_k} dy}} = \int_0^1 \sqrt{f_{\eta_{10},\eta_{11}} f_{\gamma_{10},\gamma_{11}} dy} \leq 1.$$

□

5.3.2 Kullback-Leibler divergence

The distance between Dirichlet processes could be measured by means of other indices, like the coefficients of divergence. In particular we consider now the Kullback-Leibler divergence and we show that the results are not very different from the previous ones. Such an index is not a proper distance because it is not symmetric, but it can be useful when interested in measuring distances of measures from a given one.

Definition 5.3.2 *Given the Dirichlet processes* P *and* Q *on* $(\mathcal{X},\mathcal{A})$*, their distance is given by* $d_{KL}(P,Q) = \sup_{A\in\mathcal{A}} d(P(A),Q(A))$*, where* $d(X,Y)$ *denotes the Kullback-Leibler divergence between two random variables whose distributions have densities* p *and* q *w.r.t. a dominating measure* μ*, i.e.*

$$d(X,Y) = \int p \log p/q d\mu.$$

The quantity $d(P(A),Q(A)) = Y(\eta_1(A),\eta_2(A))$ equals

$$\int_0^1 f_{\eta_1(A)}(y) \log \frac{f_{\eta_1(A)}(y)}{f_{\eta_2(A)}(y)} dy =$$

Here, $Y(x,(1-\varepsilon)x) = \dfrac{\Gamma((1-\varepsilon/2)x)\Gamma(\beta-(1-\varepsilon/2)x)}{\sqrt{\Gamma(x)\Gamma(\beta-x)\Gamma((1-\varepsilon)x)\Gamma(\beta-(1-\varepsilon)x)}}$, so that $\lim_{x\to 0}\Gamma(x)=\infty$ implies that $\lim_{x\to\beta} Y(x,(1-\varepsilon)x)=0$ and therefore there exists Q such that $d_H(P_0,Q)=\sqrt{2}$.

Example 5.3.4 Let η be in the total variation neighbourhood of η_0, which corresponds to

$$g(x) = \begin{cases} 0 & 0 \le x \le \varepsilon \\ x-\varepsilon & \varepsilon < x \le \beta \end{cases}$$

Given $0 < x < \varepsilon$, then it follows that $\lim_{t\to 0} Y(x,t) = 0$ implies that there exists Q such that $d_H(P_0,Q)=\sqrt{2}$.

The results are not surprising because the maximum distance is achieved by considering a probability measure Q which is not absolutely continuous w.r.t. P_0, as shown by $g(1) < \beta$ (see Fortini and Ruggeri [38], for details). A lesser distance is obtained when considering the class of all probability measures whose c.f.'s are not below the one described in Example 5.3.1; the same distances in Table 5.1 are now the maximum distances in the class of the Dirichlet processes.

Finally, it is worth mentioning the following result about the distance between processes when considering random vectors $P(A_1),\ldots,P(A_k)$ and $Q(A_1),\ldots,Q(A_k)$, which shows that the distance between processes increases as we consider finer partitions. Let $f_{\eta_1,\ldots,\eta_k}$ be the density of a Dirichlet distributed random variable with parameters $(\eta_1,\ldots,\eta_k)$. Let $\{B_1,\ldots,B_k\}$ be a measurable partition of $\mathcal{A}$ and $\{B_{10},B_{11}\}$ be a measurable partition of B_1; it follows that $P(B_1),\ldots,P(B_k)$ and $Q(B_1),\ldots,Q(B_k)$ are Dirichlet distributed with densities $f_{\eta_1,\ldots,\eta_k}$ and $f_{\gamma_1,\ldots,\gamma_k}$, respectively.

the c.f. of η_2 w.r.t. η_1 is given by $\varphi(x) = x + (1-x)\log(1-x)$, and numerically shown that $d_H(P,Q) \approx \sqrt{2}$. The result is not surprising since the c.f. tells us that there exist subsets whose probability is very small under η_1 and very large under η_2.

The most interesting application is about classes Π of Dirichlet processes Q determined by their parameters η being in a class Γ. Let η_0 be the parameter of a baseline Dirichlet process P_0. Let Λ the class of all c.f.'s of the parameters $\eta \in \Gamma$ w.r.t. η_0. The following Theorem simplifies the search of the distance between the process P_0 and the other processes Q. Its proof is omitted because it is very similar to the proof of Theorem 5.3.1.

Theorem 5.3.2 *For any* $x \in [0,1]$*, define* $\hat{\varphi}(x) = \inf_{\varphi \in \Lambda} \varphi(x)$*. Then it follows that*

$$\sup_{Q \in \Pi} d_H(P_0, Q) = \sqrt{2}\left(1 - \inf_{0 \le x \le \beta_1} Y(x, \hat{\varphi}(x))\right)^{1/2}.$$

The class Γ could be one of those described by Fortini and Ruggeri [37], e.g. particular cases of ε–contaminations, total variation neighbourhood, density ratio and density bounded, where neighbourhoods of a given probability measure η_0 were defined by considering all the probability measures whose c.f. w.r.t. η_0 was not below a given monotone nondecreasing, continuous, convex function $g(x)$ with $g(0) = 0$ and $g(1) \le 1$. It is even possible to consider measures with finite mass β, provided that $g(\beta) \le \beta$.

Example 5.3.3 Let η be in the ε–contaminated class $\Gamma_\varepsilon = \{\eta : \eta = (1-\varepsilon)\eta_0 + \varepsilon\gamma, \gamma \in \mathcal{M}\}$, where $\mathcal{M}$ is the class of all measures with mass β. Such a class is a neighbourhood of η_0 described by Fortini and Ruggeri [37] by taking $g(x) = (1-\varepsilon)x$.

Table 5.1: Hellinger distance

γ	.0001	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.
Distance	1.40	.98	.79	.65	.53	.42	.33	.24	.16	.08	0.

takes its minimum value at either $\inf_{A \in \mathcal{A}_x} \eta_2(A)$ or $\sup_{A \in \mathcal{A}_x} \eta_2(A)$, i.e., as discussed in Appendix B, at either $\varphi(x)$ or $\beta_2 - \varphi(\beta_1 - x)$. The same argument can be repeated for any $x \in [0, \beta_1]$. □

Therefore the distance between two processes is found by computing the c.f. $\varphi(x)$ and then minimising the function $Y(x, \varphi(x))$, as in the following examples.

Example 5.3.1 Let P and Q be Dirichlet processes on $(\Re, \mathcal{B})$ whose parameters have densities η_1 and η_2 such that $\eta_2(x) = \gamma\eta_1(x)$ on a subset A with measure 0.5 under η_1 and $\eta_2(x) = (2-\gamma)\eta_1(x)$ on A^C, with $0 \leq \gamma \leq 1$ and $\eta_1(\Re) = 1 = \eta_2(\Re)$. Therefore the c.f. of η_2 w.r.t. η_1 is given by

$$\varphi(x) = \begin{cases} \gamma x & 0 \leq x \leq 0.5 \\ (2-\gamma)x + \gamma - 1 & 0.5 \leq x < 1 \end{cases}$$

By proving that $\mathcal{E}_x L = \Psi(x) - \Psi(\beta)$ and assuming for simplicity $\beta = 1$, then, from the well-known result $\Psi(x) - \Psi(1-x) = -\pi \cot \pi x$, and some, not trivial, algebraic manipulations, it follows that $\frac{\partial Y(x, \varphi(x))}{\partial x}$ is negative (positive) for $x \leq .5$ $(x > .5)$, so that $Y(x, \varphi(x))$ achieves its minimum value at $x = .5$. Finally, it can be shown that $\lim_{\gamma \to 0} Y(x, \varphi(x)) = 0$ so that $d_H(P, Q) = \sqrt{2}$.

Distances are computed for different values of γ and shown in Table 5.1.

Example 5.3.2 Let P and Q be Dirichlet processes whose parameters η_1 and η_2 have distributions $\mathcal{E}(1)$ and $\mathcal{G}(2, 1)$, respectively. It can be proved that

$$\begin{aligned}T_x(\eta_2(A)) &= \Psi((\eta_1(A)+\eta_2(A))/2) - \Psi((\beta_1+\beta_2)/2 - (\eta_1(A)+\eta_2(A))/2) \\ &-\Psi(\eta_2(A)) + \Psi(\beta_2 - \eta_2(A)).\end{aligned}$$

It can be shown that $\lim_{\eta_2(A)\to 0} \frac{\partial Z_x}{\partial \eta_2(A)} = +\infty$ and $\lim_{\eta_2(A)\to \beta_2} \frac{\partial Z_x}{\partial \eta_2(A)} = -\infty$.

From Abramovitz and Stegun [1], p.259-260, it follows that

$$\Psi(z) = \int_0^\infty \left[\frac{e^{-t}}{t} - \frac{e^{-zt}}{1-e^{-t}}\right] dt \text{ and } \Psi'(z) = \int_0^\infty \frac{te^{-zt}}{1-e^{-t}}dt,$$

so that

$$\Psi((\eta_1(A)+\eta_2(A))/2) - \Psi(\eta_2(A)) = \int_0^\infty \frac{e^{-\eta_2(A)t}\left\{1-e^{-(\eta_1(A)-\eta_2(A))/2}\right\}}{1-e^{-t}}dt$$

and

$$\Psi(\beta_2 - \eta_2(A)) - \Psi((\beta_1+\beta_2)/2 - (\eta_1(A)+\eta_2(A))/2) =$$

$$= \int_0^\infty \frac{e^{-(\beta_2-\eta_2(A))t}\left\{e^{-((\beta_1-\eta_1(A))/2-(\beta_2-\eta_2(A))/2)t} - 1\right\}}{1-e^{-t}}dt.$$

Looking for the quantities within brackets to be positive for all t, it follows that $\frac{\partial Z_x}{\partial \eta_2(A)}$ is positive for $\eta_2(A) < \beta_2 - \beta_1 + \eta_1(A)$ and negative for $\eta_2(A) > \eta_1(A)$. When $\beta_2 - \beta_1 + \eta_1(A) \leq \eta_2(A) \leq \eta_1(A)$, then it can be shown that

$$\begin{aligned}\frac{\partial T_x}{\partial \eta_2(A)} &= \left\{1/2\Psi'((\eta_1(A)+\eta_2(A))/2) - \Psi'(\eta_2(A))\right\} + \\ &+ \left\{1/2\Psi'((\beta_1+\beta_2)/2 - (\eta_1(A)+\eta_2(A))/2) - \Psi'(\beta_2 - \eta_2(A))\right\}\end{aligned}$$

is negative.

Therefore $\frac{\partial Z_x}{\partial \eta_2(A)} = 0$ at a unique point, which is the unique maximum for $Z_x(\eta_2(A))$ (e.g. $\eta_2(A) = x$ if $\beta_1 = \beta_2$) and, since Z_x is a continuous function, it where

$$\begin{aligned} Y(\eta_1, \eta_2) &= \int_0^1 \sqrt{f_{\eta_1,\beta_1}(y) f_{\eta_2,\beta_2}(y)}dy \\ &= \frac{\Gamma((\eta_1+\eta_2)/2)\Gamma((\beta_1+\beta_2)/2-(\eta_1+\eta_2)/2)}{\sqrt{\Gamma(\eta_1)\Gamma(\beta_1-\eta_1)\Gamma(\eta_2)\Gamma(\beta_2-\eta_2)}} \cdot \frac{\sqrt{\Gamma(\beta_1)\Gamma(\beta_2)}}{\Gamma((\beta_1+\beta_2)/2)}. \end{aligned}$$

The distance $d_H(P, Q)$ can be expressed by means of the concentration function (c.f.), defined by Cifarelli and Regazzini [16]. The main properties of the c.f. are described in Appendix B, along with its extension to the comparison between finite, positive measures with different total mass.

Without loss of generality, suppose that $\beta_2 \leq \beta_1$. We consider the c.f. φ_1 of η_2 w.r.t. η_1, but we should notice that we would have got the same result by considering the c.f. φ_2 of η_1 w.r.t. η_2. As discussed in Appendix B, it is false, in general, that $\varphi_1(t) = \varphi_2(t), \forall t \in [0, 1]$, but here it is possible to prove that $Y(x, \varphi_1(x)) = Y(\varphi_1(x), \beta_2 - \varphi_2(\beta_1 - \varphi_1(x)))$. Furthermore, it can be proved that $Y(x, \beta - \varphi(\beta - x)) = Y(\beta - x, \varphi(\beta - x))$, so that it is sufficient to consider $Y(x, \varphi(x))$.

Theorem 5.3.1 $d_H(P, Q) = \sqrt{2}\left(1 - \inf_{0\leq x\leq\beta_1} Y(x, \varphi(x))\right)^{1/2}$, *where* φ *is the c.f. of* η_2 *w.r.t.* η_1.

PROOF. Consider $\mathcal{A}_x = \{A \in \mathcal{A} : \eta_1(A) = x\}$, $0 \leq x \leq \beta_1$; we want to find $A \in \mathcal{A}_x$ which minimises $Y(x, \eta_2(A))$. Take

$$Z_x(\eta_2(A)) = Y(x, \eta_2(A)) \frac{\sqrt{\Gamma(\eta_1(A))\Gamma(\beta_1 - \eta_1(A))}\Gamma((\beta_1 + \beta_2)/2)}{\sqrt{\Gamma(\beta_1)\Gamma(\beta_2)}}.$$

It follows that $\dfrac{\partial Z_x}{\partial \eta_2(A)} = Z_x(\eta_2(A))/2 \cdot T_x(\eta_2(A))$, where $\Psi(x) = \dfrac{\partial \log(\Gamma(x))}{\partial x}$ and of $\mathcal{A}$ (see Definition 2.2.1.).

5.3.1 Hellinger distance

Definition 5.3.1 *Given the Dirichlet processes* P *and* Q *on* $(\mathcal{X}, \mathcal{A})$*, their distance is given by* $d_H(P,Q) = \sup_{A \in \mathcal{A}} d(P(A), Q(A))$*, where* $d(X,Y)$ *denotes the Hellinger distance between two random variables whose distributions have densities* p *and* q *w.r.t. a dominating measure* μ*, i.e.*

$$d(X,Y) = \left\{ \int (\sqrt{p} - \sqrt{q})^2 \, d\mu \right\}^{1/2}.$$

d_H is actually a distance; in fact, it is symmetric and nonnegative. Besides, $d_H(P,Q) = 0$ if and only if the processes are the same, i.e. they have the same parameters a.e.. From Definition 2.2.1., it follows that, for any $A \in \mathcal{A}$, $P(A)$ and $Q(A)$ are Beta distributed, with parameters $(\eta_1(A), \beta - \eta_1(A))$ and $(\eta_2(A), \beta - \eta_2(A))$. The condition $d_H(P,Q) = 0$ implies $d(P(A), Q(A)) = 0$ for all $A \in \mathcal{A}$, so that the two Beta distributions coincide and $\eta_1(A) = \eta_2(A)$ for all $A \in \mathcal{A}$ and the two Dirichlet processes have the same parameter η. Vice versa, two Dirichlet processes with the same parameter are such that $d_H(P,Q) = 0$. Finally, given the processes P, Q and R, the triangle inequality is proved by

$$d_H(P,Q) \leq \sup_{A \in \mathcal{A}} \{d(P(A), R(A)) + d(R(A), Q(A))\} \leq d_H(P,R) + d_H(R,Q).$$

Let $f_{\eta,\beta}$ be the density of a Beta distributed random variable with parameters $(\eta, \beta - \eta)$. When evident (e.g. $\eta = \eta(A)$ and $\beta = \eta(\mathcal{X})$), then the subscript β will be omitted. Given two measures η_1 and η_2 such that $\eta_1(\mathcal{X}) = \beta_1$ and $\eta_2(\mathcal{X}) = \beta_2$ respectively, it can be easily shown that

$$d(P(A), Q(A)) = \sqrt{2} \, (1 - Y(\eta_1(A), \eta_2(A)))^{1/2},$$

and density ratio classes, might be considered as well.

Therefore, as the measure α belongs to a class Ψ, we have a family of Dirichlet processes, which play the same role as the prior distributions in the robust parametric approach, and we want to look at the behaviour of some quantity of interest and compute its range, as α varies in Ψ. Here, we are particularly interested in the distance between Dirichlet processes, in the probability of some subspace of the space of all probabilities on $(\mathcal{X}, \mathcal{A})$, defined by means of either the probabilities of some measurable spaces in $\mathcal{A}$ or the expectations of some measurable function on $(\mathcal{X}, \mathcal{A})$; in the probability of such probabilities or expectations and, finally, in some Bayes estimators, e.g. of a random distribution.

5.3 Distance between Dirichlet processes

As the parameter η of a Dirichlet process varies in a class, a family of Dirichlet processes is obtained and a sensitivity analysis could be performed about how "far" these processes are from one another. The idea of measuring the distance between stochastic processes is, of course, not new in literature. We simply mention Vajda [65] where Renyi distances and Kullback-Leibler divergence, both based on Hellinger integrals, were used to evaluate the distances between distributions of regular Markov processes. Here we consider, as a distance between two Dirichlet processes, both the maximum Hellinger distance and the maximum Kullback-Leibler divergence between the distributions, under the two Dirichlet processes, of the probability of any subset $A \in \mathcal{A}$. The interest in such distributions follows, quite naturally, from the definition of Dirichlet process, in terms of distribution of random probabilities of partitions shrinking towards a given prior probability measure, like in Ruggeri and Wasserman [63] and Ruggeri [60].

The author has analysed elsewhere some classes of priors (see Ruggeri [55], [56], [58], [59], Betrò, Męczarski and Ruggeri [11], Fortini and Ruggeri [35], [36], [37], [38], Carota and Ruggeri [15]) and classes of procedures (Betrò and Ruggeri [12] and Ríos Insua, Ruggeri and B.Vidakovic [51]).

5.2 Nonparametric Bayesian robustness

The parameter η of a Dirichlet process is arbitrarily chosen by the statistician or the practitioner according to his/her own belief. In particular, α is often chosen as the distribution function that the scientist believes, a priori, to be the best in describing the behaviour of the phenomenon that he/she is analysing by means of Bayesian nonparametric inference; such a choice follows from Proposition 2.2.1 because α is the expected value of the random distribution function governing the phenomenon. Besides, β denotes the "strength of belief" in the prior measure η, because the posterior Bayes estimator of the random distribution, under squared loss, is a mixture of η and the empirical distribution, with greater weight for η for greater β's (see Ferguson [30].)

Because of the arbitrariness in specifying α and β, it may be possible that they cannot be specified exactly, e.g. the joint study of incomes and political opinions in a population could be performed by a Bayesian nonparametric approach, eliciting only the marginals of the two quantities, by using the data about tax returns and poll results, discounted for the possible biases. In such a case, the probability measure α would be in a Fréchet class, like broadly considered in this work. Other cases, like ε–contaminations, density bounded posterior distribution for θ.

In parametric Bayesian analysis, a single prior probability measure is usually chosen from among those compatible with the prior knowledge, but the choice could heavily affect the inference about the parameters.

Robust Bayesian methodology usually models the uncertainty in the prior by using a class Γ of probability measures on $(\Theta, \mathcal{F})$, and studies the ranges of posterior functions of interest (e.g. probability of sets, mean, variance and loss function) as the prior varies in Γ. A small range means that the results are not meaningfully different, ensuring robustness with respect to (w.r.t.) the prior probability measure. Depending on the problem, different classes Γ of probability measures may be considered. Ideally, the class Γ should be easily specified, large enough to contain any plausible prior and small enough to exclude unnatural priors, and involve relatively simple computations.

The study of bounds on posterior quantities over a class of priors is usually referred to as global sensitivity analysis. Such an approach has received much attention; see Berger ([4], [5], [6], [7]) and Wasserman ([67]) and the references contained therein.

Recently, Bayesian robustness studies have also considered local sensitivity analyses, in which posterior effects of infinitesimal departures from an elicited prior are studied. Extending work by Diaconis and Freedman [27], Ruggeri and Wasserman [62] analysed some ε–contamination classes of priors, calculating the norm of the Fréchet derivative of the posterior expectation of a function with respect to the prior over the class of all signed measures with zero mass. Otherwise, the analysis is performed considering the derivatives, with respect to a parameter ε, of the bounds on the posterior expectation over classes Γ_ε

Chapter 5

Bayesian robustness

In this section we present a new, nonparametric, approach to Bayesian robustness. Whereas many studies in Bayesian robustness have dealt with a parametric sampling distribution, considering classes of prior distributions on the parameters, here we assume that the sampling distribution comes from a Dirichlet process with a parameter $\eta = \beta\alpha$, with $\beta > 0$ and α being a probability measure, specified with uncertainty.

5.1 Parametric Bayesian robustness

In a parametric framework, the distribution of a random variable X, defined on a dominated statistical space denoted by $(\mathcal{X}, \mathcal{F}_{\mathcal{X}}, \{P_\theta, \theta \in (\Theta, \mathcal{F})\})$, with density $f(x|\theta)$, depends on a parameter θ, considered as a random variable, for which a prior distribution function is defined on the parameter space $(\Theta, \mathcal{F})$, with Θ being a subset of a finite-dimensional Euclidean space. Given an observation x of X, the experimental evidence about θ will be expressed by the likelihood function $l_x(\theta) = f(x|\theta)$, which we assume $\mathcal{F}_{\mathcal{X}} \times \mathcal{F}$-measurable. Combining likelihood and prior distribution, it is well known that we get the so that $\tilde{\alpha}(z_i) \neq \tilde{H}(z_i)$. Given $i > i_0$, suppose that $\tilde{\alpha}(z_i) > \tilde{H}(z_i)$, then it leads to the false statement $\tilde{\alpha}(z_i) > \hat{M}(z_i)$; similarly, $\tilde{\alpha}(z_i) < \tilde{H}(z_i)$ implies $\tilde{\alpha}(z_i) < \hat{W}(z_i)$, which is impossible. $\square$

4.6 Distance from a given distribution

In some problems (e.g. the earthquake model in Chapter 6) we might be interested in measuring the distance between a touchstone distribution $\tilde{H}$ and an estimate (usually Bayes) of a random distribution H from a Dirichlet process $\mathcal{D}(\eta)$. Allowing for uncertainty in the measure η, the value of such distance spans a range, which tells us how far, and how close, the estimates are to $\tilde{H}$.

In particular, we consider $\alpha \in \Pi(\hat{W}, \hat{M})$ and the Bayes estimator $\mathcal{E}H = \alpha$ of the distribution H. A very tractable distance, which takes into account the fact that $\tilde{H}$ could come from an absolutely continuous probability measure (with respect to Lebesgue measure) whereas the Bayes estimators cannot (see Ferguson [30], is given by the supremum norm, i.e.

$$d(\alpha, \tilde{H}) = \sup_{z \in \Re^2} |\alpha(z) - \tilde{H}(z)|.$$

The upper bound on the distance is given by the following result (the lower bound is of no interest).

Theorem 4.6.1

$$\sup_{\alpha \in \Pi(\hat{W}, \hat{M})} d(\alpha, \tilde{H}) = \max\{d(\hat{W}, \tilde{H}), d(\hat{M}, \tilde{H})\}.$$

Proof. The result is proved by contradiction. Suppose that the supremum $\overline{d}$ is achieved at $\tilde{\alpha} \neq \hat{W}, \hat{M}$, then there exists a sequence of points $\{z_i\}_{i=1}^{\infty}$ such that $|\tilde{\alpha}(z_i) - \tilde{H}(z_i)| \uparrow \overline{d}$ (note that we need the sequence because we cannot assure that the supremum is achieved at a point in the domain of Z). There exists i_0 such that for every $i > i_0$, then

$$|\tilde{\alpha}(z_i) - \tilde{H}(z_i)| > \max\{|\hat{M}(z_i) - \tilde{H}(z_i)|, |\hat{W}(z_i) - \tilde{H}(z_i)|\},$$

$$\alpha(B_i) = \frac{\beta \Delta F_i + \sum_{i=1}^n \delta_{Z_i}(B)}{\beta + n} \text{ and } \alpha(A) = \frac{\beta \Delta G_i + \sum_{i=1}^n \delta_{Z_i}(A)}{\beta + n}$$

are constant in $\Pi(\hat{W}, \hat{M})$, whereas

$$\sup_{\alpha \in \Pi(\hat{W}, \hat{M})} \alpha(A \cap B) = \frac{\beta \min\{\Delta F_i, \Delta G\} + \sum_{i=1}^n \delta_{Z_i}(A \cap B)}{\beta + n},$$

and

$$\inf_{\alpha \in \Pi(\hat{W}, \hat{M})} \alpha(A \cap B) = \frac{\beta \max\{\Delta F_i + \Delta G - 1, 0\} + \sum_{i=1}^n \delta_{Z_i}(A \cap B)}{\beta + n}.$$

Therefore (4.5.1) is an immediate consequence.

We show that $\Pi(\hat{W}, \hat{M})$ is dilation prone by showing that there exists A and $\mathcal{B}$ such that $\mathcal{B}$ dilates A.

Given a sample of size N, we can always find a subset A such that it contains no observation and such that $0 < \Delta G < \min\{\Delta F_1, \Delta F_2\} < 1$, where $\mathcal{B} = \{B_1, B_2\}$ is such that $\Delta F_1 = \Delta F_2 = 0.5$. It can be easily shown that both inequalities in (4.5.1.) are satisfied.

Finally, it can be shown that not all the partitions dilate A. Suppose that both A and an element $B_i \in \mathcal{B}$ contain no observation, that $\Delta F_i + \Delta G > 1$ and the size n is sufficiently large, thus the left hand inequality in (4.5.1) may be false. $\square$

Note that we could have proved the result by using, at a certain point, Theorem 2.3 in Seidenfeld and Wasserman [64], but we think that the direct proof is much more understandable.

We consider any $A = \Re \times (y_0, y_1]$, *with* $0 < \Delta G = G(y_1) - G(y_0) < 1$, *and any partition* $\mathcal{B}$ *with elements* $B_i = (x_{i-1}, x_i] \times \Re$, $-\infty = x_0 < x_1 < \ldots < x_n = \infty$, *such* $0 < \Delta F_i = F(x_i) - F(x_{i-1}) < 1$.

Given the class $\Gamma(F, G)$, *then any such partition* $\mathcal{B}$ *strictly dilates any such* A. *Given the class* $\Pi(\hat{W}, \hat{M})$, *then the partition* $\mathcal{B}$ *dilates* A *if and only if, for any* $i = 1, \ldots, n$,

$$\max\{\Delta F_i + \Delta G - 1, 0\} \leq$$

$$\frac{\{\beta \Delta F_i + \sum_{i=1}^{n} \delta_{Z_i}(B)\}\{\beta \Delta G + \sum_{i=1}^{n} \delta_{Z_i}(A)\} - (\beta + n) \sum_{i=1}^{n} \delta_{Z_i}(A \cap B)}{\beta(\beta + n)} \tag{5.4.1}$$

$$\leq \min\{\Delta F_i, \Delta G\}.$$

Proof. Since $\alpha(B_i) = \Delta F_i$ is constant in $\Gamma(F, G)$, thus upper and lower conditional probability are found by looking at the bounds on $P(A \cap B)$. It can be shown that

$$\sup_{\alpha \in \Gamma(F,G)} \alpha(A \cap B) = \min\{\Delta F_i, \Delta G\}$$

and

$$\inf_{\alpha \in \Gamma(F,G)} \alpha(A \cap B) = \max\{\Delta F_i + \Delta G - 1, 0\}.$$

It follows that upper and lower conditional probabilities are strictly greater and smaller, respectively, than $P(A) = \Delta G$ for all $\alpha \in \Gamma(F, G)$, since

$$\frac{\max\{\Delta F_i + \Delta G - 1, 0\}}{\Delta F_i} < \Delta G < \frac{\min\{\Delta F_i, \Delta G\}}{\Delta F_i}.$$

Therefore, any such partition $\mathcal{B}$ dilates A and so $\Gamma(F, G)$ is dilation prone.

Considering the posterior class, then it follows that whereas taking $H_x = F_x - \delta$, $0 \leq \delta \leq F_x - W_x$, implies that the supremum $\overline{E}$ is achieved at $\delta = 0$. Then it follows that

$$\overline{E} = \frac{1 - G + \gamma_x}{1 - G + \gamma_0}.$$

If $F_x > G$, then $\mathcal{E}P(A|B)$ is maximised by taking $H_x - H_0 = \min\{F_x - F_0, G\}$ and $H_0 = \max\{G + F_0 - F_x, 0\}$; if $F_x - F_0 \leq G$, then $H_x = M_x = G$ and $H_0 = G - F_x + F_0$, so that

$$\overline{E} = \frac{1 - F_x + \gamma_x}{1 - F_x + \gamma_0};$$

finally, if $F_x - F_0 > G$, then $H_x = M_x = G$ and $H_0 = G - G = 0$, so that

$$\overline{E} = \frac{1 - F_x + \gamma_x}{1 - F_0 - G + \gamma_0}.$$

4.5 Dilation

Dilation is an interesting phenomenon recently discussed by Seidenfeld and Wasserman [64] and [68] and shortly reviewed in Appendix D. Here we consider $\mathcal{E}P(A)$ and $\mathcal{E}P(A|B)$, for both prior and posterior Dirichlet processes, when considering the prior parameter $\eta = \beta\alpha$ such that α is in the Fréchet class $\Gamma(F, G)$, so that the posterior α is in $\Pi(\hat{W}, \hat{M})$.

It is known that $\mathcal{E}P(A) = \alpha(A)$ (Proposition 2.2.1) and that $\mathcal{E}P(A|B) = \alpha(A \cap B)/\alpha(B)$ (Corollary 4.4.3).

We study now dilation for the classes of probability measures $\Gamma(F, G)$ and $\Pi(\hat{W}, \hat{M})$. For sake of simplicity, we suppose that F and G are continuous, increasing distribution functions.

Theorem 4.5.1 *The classes $\Gamma(F, G)$ and $\Pi(\hat{W}, \hat{M})$ are dilation prone.*

If the prior Frèchet bound W_x equals 0, then it follows that

$$\underline{E} = \frac{1 - F_x - G + \gamma_x}{1 - F_0 - G + \gamma_0} < 1,$$

by taking $H_x = W_x = W_0 = H_0 = 0$.

If $W_x > 0$ and $W_0 = 0$, then

$$\underline{E} = \frac{\gamma_x}{1 - F_0 - G + \gamma_0} < 1,$$

by taking $H_x = W_x$ and $W_0 = H_0 = 0$.

IF $W_x > 0$ and $W_0 > 0$, then different cases are possible, depending on F_0 and G. The underlying idea is to have H_0 as large as possible and $H_x - H_0$ as small as possible. It can be seen that $F_0 \geq G$ implies that

$$\underline{E} = \frac{1 - F_x + \gamma_x}{1 - F_0 + \gamma_0},$$

by taking $H_x = H_0 = G$.

If $F_0 < G$, then take $H_0 = F_0$ and $H_x = F_0$ if $F_x + G - 1 \leq F_0$, and $H_x = W_x$ otherwise. In the former case , it follows that

$$\underline{E} = \frac{1 - F_x - G + F_0 + \gamma_x}{1 - G + \gamma_0},$$

whereas in the latter,

$$\underline{E} = \frac{\gamma_x}{1 - G + \gamma_0} < 1.$$

To get the supremum $\overline{E}$ of $\mathcal{E}P(A|B)$, we need to find the infimum of C and we will show that it happens by maximising H_x and $H_x - H_0$.

If $F_x \leq G$, then $\mathcal{E}P(A|B)$ is maximised by taking $H_x = M_x = F_x$ and $H_0 = M_0 = F_0$; in fact, $F_x - F_0$ is the maximum value achievable by $H_x - H_0$, Given the probability measure α^*, we use the following simplified, evident notations: the joint distribution function at (x^-, y^-) is $\dfrac{\beta H_x + Z_x}{\beta + n}$, while the marginals are $\dfrac{\beta F_x + X_x}{\beta + n}$ and $\dfrac{\beta G + Y}{\beta + n}$. The subscript 0, in lieu of x, denotes the same distributions at (x_0^-, y^-).

We suppose that $\alpha([x_0, +\infty) \times ([y, +\infty)) > 0$ and, to avoid $\mathcal{E}P(A|B) = 1$ for all α, that $F_x - F_0 > 0$. Therefore $\mathcal{E}P(A|B)$ equals

$$\frac{\alpha([x, +\infty) \times [y, +\infty))}{\alpha([x_0, +\infty) \times ([y, +\infty)))} =$$
$$= \frac{1 - (\beta F_x + X_x)/(\beta + n) - (\beta G + Y)/(\beta + n) + (\beta H_x + Z_x)/(\beta + n)}{1 - (\beta F_0 + X_0)/(\beta + n) - (\beta G + Y)/(\beta + n) + (\beta H_0 + Z_0)/(\beta + n)}$$
$$= \frac{1 - F_x - G + H_x + \gamma_x}{1 - F_0 - G + H_0 + \gamma_0}$$
$$= (1 + C)^{-1},$$

where

$$C = \frac{(F_x - F_0) - (H_x - H_0) - (\gamma_x - \gamma_0)}{1 - F_x - G + H_x + \gamma_x},$$

and $\beta\gamma_x$ ($\beta\gamma_0$) equals the number of observations (s, t) with $s \geq x$ ($s \geq x_0$) and $t \geq y$. Note that the prior Dirichlet process can be considered too, provided that we take $\gamma_x = \gamma_0 = 0$.

We find the bounds only for $C = \dfrac{\alpha([x_0, x) \times [y, +\infty))}{\alpha([x, +\infty) \times ([y, +\infty)))}$. Actually, being γ_x and γ_0 constants, the problem considers only prior distributions.

To get the infimum $\underline{E}$ of $\mathcal{E}P(A|B)$, we need to find the supremum of C and we will show that it happens by minimising H_x and $H_x - H_0$.

It follows that T_k is independent of $(T_1, \dots, T_{k-1})$, so that we can integrate it out to get the density of $(T_1, \dots, T_{k-1})$, which corresponds to a Dirichlet distribution with parameters $(\eta_1, \dots, , \eta_k)$. □

Corollary 4.4.1 *Let P be a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter η, and let $A \in \mathcal{A}$. Then the conditional distribution of $(1/M)P$, given $P(A) = M$, and the distribution of the conditional probability $P(\cdot|A)$ are Dirichlet process with the same parameter.*

Proof. It follows from Theorem 2.2.4 and Theorem 4.4.2. □

The distribution of the random conditional probability of a subset is an immediate consequence of Theorem 4.4.2.

Corollary 4.4.2 *Let P be a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter η, then*

$$P(A|B) \sim \mathcal{B}(\eta(A \cap B), \eta(A^C \cap B)),$$

where $\mathcal{B}$ denotes the Beta distribution and $A, B \in \mathcal{A}$ are such that $\eta(A \cap B)\eta(A^C \cap B)\eta(B^C) \neq 0$.

Corollary 4.4.3 *Under the same assumptions as in Theorems 4.4.1 and 4.4.2, it follows that $\mathcal{P}(Z \in A|Z \in B) = \mathcal{E}P(A|B) = \alpha(A \cap B)/\alpha(B)$.*

Proof. It follows immediately by looking at the expected value of a random variable distributed according to a Beta distribution with parameters $(\eta(A \cap B), \eta(A^C \cap B))$. □

We will now compute the range of $\mathcal{E}P(A|B)$ as α^* varies in $\Psi(\hat{W}, \hat{M})$, when, given $x_0 \leq x$, $B = [x_0, +\infty) \times [y, +\infty)$ and $A = [x, \infty) \times ([y, +\infty)$.

The Bayes estimator of $P(A|B)$ under squared loss function is given by $\mathcal{E}P(A|B) = \dfrac{\alpha^*(A \cap B)}{\alpha^*(B)}$.

Theorem 4.4.1 *Let P be a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter η, and let Z be a sample of size 1 from P. Then for $A, B \in \mathcal{A}$, $\mathcal{P}(Z \in A|Z \in B) = \alpha(A \cap B)/\alpha(B)$.*

PROOF. Applying Proposition 2.2.2, it follows that

$$\mathcal{P}(Z \in A|Z \in B) = \frac{\mathcal{P}(Z \in A \cap B)}{\mathcal{P}(Z \in B)} = \frac{\alpha(A \cap B)}{\alpha(B)}.$$

□

Theorem 4.4.2 *Let P be a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter η, and let $A \in \mathcal{A}$. Then the distribution of the conditional probability $P(\cdot|A)$ is a Dirichlet process on $(A, \mathcal{A} \cap A)$ with parameter η restricted to A. That is, if $A_1, \ldots, A_k$ is any measurable partition of A, then the distribution of $(P(A_1|A), \ldots, P(A_k|A))$ is a Dirichlet distribution with parameter $(\eta(A_1), \ldots, \eta(A_k))$.*

PROOF. To simplify the notation, define $\eta(A_i \cap A) = \eta_i$, for $i = 1, \ldots, k$, and $\eta(A^C) = \eta_{k+1}$, $Y_i = P(A_i \cap A)$, for $i = 1, \ldots, k$, and $Y_{k+1} = P(A^C)$. From Definition 2.2.1, it follows that $(Y_1, \ldots, Y_{k+1}) \sim \mathcal{D}(\eta_1, \ldots, \eta_{k+1})$.

Consider the random variables $T_k = \sum_{i=1}^{k} Y_i$ and $T_i = \frac{Y_i}{\sum_{i=1}^k Y_i}$, for $i = 1, \ldots, k-1$, so that $0 \leq T_i \leq 1$, for $i = 1, \ldots, k$, and $Y_i = T_i T_k$ for $i = 1, \ldots, k-1$, and $Y_k = T_k(1 - \sum_{i=1}^{k-1} T_i)$. Since the Jacobian in the change of variables is equal to T_k^{k-1}, it follows that the joint density of $T_1, \ldots, T_k$ is given by

$$\begin{aligned} f_{T_1,\ldots,T_k}(t_1, \ldots, t_k) &= K t_k^{k-1} \textstyle\prod_{i=1}^{k-1} (t_k t_i)^{\eta_i - 1} \{t_k(1 - \sum_{i=1}^{k-1} t_i)\}^{\eta_k - 1} \cdot \\ &\cdot \{1 - t_k \textstyle\sum_{i=1}^{k-1} t_i - t_k(1 - \sum_{i=1}^{k-1} t_i)\}^{\eta_{k+1} - 1} \\ &= K t_k^{\sum_{i=1}^k \eta_i - 1} (1 - t_k)^{\eta_{k+1} - 1} \textstyle\prod_{i=1}^{k-1} t_i^{\eta_i - 1} (1 - \sum_{i=1}^{k-1} t_i)^{\eta_k - 1}. \end{aligned}$$

and from Theorem 4.1.8 which shows that

$$\lim_{n\to+\infty} \hat{M}(x,y) = \lim_{n\to+\infty} \hat{W}(x,y) = \tilde{H}(x,y).$$

Finally observe that $\lim_{n\to+\infty} \inf_{F\in\mathcal{F}} \sup_{\alpha^*\in\Psi(\hat{W},\hat{M})} r(\alpha^*,\hat{F})$ equals

$$\int_{\Re^2} \{M(x,y) - W(x,y)\}^2 dR(x,y) \lim_{n\to+\infty} \left\{\frac{\beta}{2(\beta+n)}\right\}^2 = 0.$$

□

Remark 4.3.3 The posterior regret action $\hat{F}$ is therefore a consistent estimator of $\tilde{H}$ in the sense that it converges, pointwise, to $\tilde{H}(x,y)$, for all x in $\Re^2$, as n goes to $+\infty$.

4.4 Conditional probability

Many statistical applications involving two random quantities X and Y require the study of one of them depending on the behaviour of the other, as expressed by the conditional probability $P(X \in A|Y \in B)$, where A and B are measurable subsets of the spaces in which X and Y are, respectively, defined. We suppose that the random joint distribution function of $Z = (X,Y)$ is given by a Dirichlet process with parameter η. Consider the subsets $A_1 = A \times \Re$ and $B_1 = \Re \times B$; we will prove that the conditional probability $P(X \in A|Y \in B) = P(Z \in A_1|Z \in B_1)$ is given by a Dirichlet process, too.

Some of the results that we are going to prove, which have not been published in any paper we know, are valid for samples from any Dirichlet process, unless differently stated.

PROOF. The convergence result for $\hat{F}$ is proved by observing that

$$\left|\frac{\hat{W}+\hat{M}}{2}-\frac{W+M}{2}\right| \leq \left|\frac{\hat{W}-W}{2}\right| + \left|\frac{\hat{M}-M}{2}\right|, \ \forall (x,y) \in \Re^2,$$

and the two terms in the right hand side go to 0, as β goes to $+\infty$, because of Theorem 4.1.7.

Finally observe that

$$\lim_{\beta\to+\infty} \inf_{F\in\mathcal{F}} \sup_{\alpha^*\in\Psi(\hat{W},\hat{M})} r(\alpha^*,\hat{F}) =$$

$$\begin{aligned} &= \int_{\Re^2} \{M(x,y)-W(x,y)\}^2 dR(x,y) \lim_{\beta\to+\infty} \left\{\frac{\beta}{2(\beta+n)}\right\}^2 \\ &= \int_{\Re^2} \{M(x,y)-W(x,y)\}^2 dR(x,y). \end{aligned}$$

□

Suppose now that β is kept fixed when the sample size n goes to infinity. Consider a sequence of observations $Z_1 = (X_1, Y_1), Z_2 = (X_2, Y_2), \ldots$ assumed to be conditionally independent and identically distributed (i.i.d.) with joint distribution $\tilde{H}(x,y)$ and marginals $\tilde{F}(x)$ and $\tilde{G}(y)$.

Corollary 4.3.2 *Using the same notation as in Theorem 4.3.2, it follows that*

$$\lim_{n\to+\infty} \hat{F}(x,y) = \tilde{H}(x,y), \ \forall (x,y) \in \Re^2,$$

and

$$\lim_{n\to+\infty} \inf_{F\in\mathcal{F}} \sup_{\alpha^*\in\Psi(\hat{W},\hat{M})} r(\alpha^*,\hat{F}) = 0.$$

PROOF. The convergence result for $\hat{F}$ is proved by observing that

$$\hat{W}(x,y) \leq \hat{F}(x,y) \leq \hat{M}(x,y), \ \forall (x,y) \in \Re^2,$$

$$
\begin{aligned}
C &= \int_{\Re^2} \{\hat{M}(x,y) - F(x,y) + F(x,y) - W^*(x,y)\}^2 dR(x,y) \\
&= A + B + 2\int_{\Re^2} \{\hat{M}(x,y) - F(x,y)\}\{F(x,y) - W^*(x,y)\} dR(x,y) \\
&\leq A + B + 2\int_{\Re^2} |\hat{M}(x,y) - F(x,y)||F(x,y) - W^*(x,y)| dR(x,y) \\
&\leq A + B + 2\sqrt{A}\sqrt{B} \\
&= (\sqrt{A} + \sqrt{B})^2 \\
&< (\sqrt{C/4} + \sqrt{C/4})^2 \\
&= C.
\end{aligned}
$$

Therefore, F^* is the optimal estimator $\hat{F}$, according to the posterior regret criterion.

The value of $\inf_{F\in\mathcal{F}} \sup_{\alpha^*\in\Psi(\hat{W},\hat{M})} r(\alpha^*, \hat{F})$ is finally obtained by simply substituting the expressions for $\hat{W}$ and $\hat{M}$.

So far, we have proved that $\hat{F}$ is the estimator in $\Psi(\hat{W}, \hat{M})$, but it is also in $\Pi(\hat{W}, \hat{M})$, since both $\hat{F}$ and α^* maximising $r(\alpha^*, F^*)$ (i.e. either $\hat{W}$ or $\hat{M}$) are in $\Pi(\hat{W}, \hat{M})$. $\square$

Remark 4.3.2. It is worth observing that $\inf_{F\in\mathcal{F}} \sup_{\alpha^*\in\Psi(\hat{W},\hat{M})} r(\alpha^*, \hat{F})$ does not depend on the values of the data Z_i, but simply on the sample size. The same is not true for $\hat{F}$ and the corresponding α^*, because they depend on the values, through $\hat{W}$ and $\hat{M}$.

Corollary 4.3.1 *Using the same notation as in Theorem 4.3.2, it follows that*

$$
\lim_{\beta\to+\infty} \hat{F}(x,y) = \frac{W(x,y) + M(x,y)}{2}, \ \forall (x,y) \in \Re^2,
$$

and

$$
\lim_{\beta\to+\infty} \inf_{F\in\mathcal{F}} \sup_{\alpha^*\in\Psi(\hat{W},\hat{M})} r(\alpha^*, \hat{F}) = \int_{\Re^2} \{M(x,y) - W(x,y)\}^2 dR(x,y).
$$

and is achieved by taking $\hat{F}(x,y) = \{\hat{W}(x,y) + \hat{M}(x,y)\}/2$.

PROOF. The function $F^* = (\hat{W} + \hat{M})/2$ is a proper distribution function, being a convex combination of two distributions, and it belongs to the class $\Psi(\hat{W}, \hat{M})$, too. We will compute the supremum, all over $\alpha^* \in \Psi(\hat{W}, \hat{M})$, of the posterior regret when estimating F_P by F^*, and we will prove that it is not greater than the maximum posterior regret when considering any other estimator for F_P.

For compactness of notation, let $C = \int_{\Re^2}\{\hat{M}(x,y) - W^*(x,y)\}^2 dR(x,y)$, $A = \int_{\Re^2}\{\hat{M}(x,y) - F(x,y)\}^2 dR(x,y)$ and $B = \int_{\Re^2}\{\hat{W}(x,y) - F(x,y)\}^2 dR(x,y)$, where F is any distribution function.

It can be easily shown that the $\sup_{\alpha^* \in \Psi(\hat{W},\hat{M})} r(\alpha^*, F^*)$ is obtained by taking α equal to either $\hat{M}$ or $\hat{W}$, so that it equals

$$\begin{aligned}\int_{\Re^2}\{\hat{M}(x,y) - F^*(x,y)\}^2 dR(x,y) &= \int_{\Re^2}\{\hat{W}(x,y) - F^*(x,y)\}^2 dR(x,y) \\ &= \frac{1}{4}\int_{\Re^2}\{\hat{M}(x,y) - W^*(x,y)\}^2 dR(x,y) \\ &= \frac{1}{4}C.\end{aligned}$$

Given any distribution F, we want to prove that $\sup_{\alpha^* \in \Psi(\hat{W},\hat{M})} r(\alpha^*, F) \geq \frac{1}{4}C$. But, $\sup_{\alpha^* \in \Psi(\hat{W},\hat{M})} r(\alpha^*, F) \geq \max\{A, B\}$, so it suffices to prove that $\max\{A, B\} \geq \frac{1}{4}C$. We will prove it by contradiction. If we suppose that $\max\{A, B\} < \frac{1}{4}C$, we get a contradiction since

□

Consider a sequence of observations $Z_1 = (X_1, Y_1), Z_2 = (X_2, Y_2), \ldots$ assumed independent and identically distributed (i.i.d.) with joint distribution $\tilde{H}(x, y)$ and marginals $\tilde{F}(x)$ and $\tilde{G}(y)$.

Remark 4.3.1 It is worth observing that $\hat{F}$ does not depend on the weight function R. From Theorem 4.1.8, it follows that $\hat{F}$ is a consistent estimator of $\tilde{H}$ in the sense that it converges, pointwise, to $\tilde{H}(x, y)$, for all x in $\Re^2$, as n goes to $+\infty$.

4.3.2 Minimax posterior regret criterion

We consider now another criterion based on the posterior regret, see related work in DasGupta and Zen [26] and Rios Insua, Ruggeri and Vidakovic [51]. The expected posterior regret in choosing $\hat{F}$ to estimate the random distribution function F_P from a Dirichlet process $P \in \mathcal{D}(\eta)$, is given by

$$r(\alpha, \hat{F}) = \mathcal{E}L(F_P, \hat{F}) - \inf_{\tilde{F} \in \mathcal{F}} \mathcal{E}L(F_P, \tilde{F}).$$

where L is the loss function.

Lemma 4.3.1 *Considering the loss function*

$$L(F_P, \tilde{F}) = \int_{\Re^2} \{F_P(x, y) - \tilde{F}(x, y)\}^2 dR(x, y),$$

it follows that $r(\alpha, \tilde{F}) = \int_{\Re^2} \{\tilde{F}(x, y) - \alpha(x, y)\}^2 dR(x, y)$.

PROOF. Since $\mathcal{E}F_P$ minimises such expected loss, as shown in Ferguson [30], it follows that $r(\alpha, \tilde{F}) = \mathcal{E}L(F_P, \tilde{F}) - L(F_P, \mathcal{E}F_P)$. □

Theorem 4.3.2 *The minimax posterior regret is*

$$\inf_{\tilde{F} \in \mathcal{F}} \sup_{\alpha^* \in \Psi(\hat{W}, \hat{M})} r(\alpha^*, \tilde{F}) = \left\{ \frac{\beta}{2(\beta + n)} \right\}^2 \int_{\Re^2} \{M(x, y) - W(x, y)\}^2 dR(x, y).$$

Proof. Given $\alpha^* \in \Psi(\hat{W}, \hat{M})$ (we are considering Ψ rather than Π), then for $P \in \mathcal{D}(\eta^*)$, the Bayes risk

$$\mathcal{E}L(F_P, \tilde{F}) = \int_{\Re^2} \mathcal{E}\{F_P(x,y) - \tilde{F}(x,y)\}^2 dR(x,y)$$

is minimised, for any (x, y), by choosing $\tilde{F}(x,y)$ which minimises $\mathcal{E}\{F_P(x,y) - \tilde{F}(x,y)\}^2$. Thus, as in Ferguson [30], the Bayes rule is $\tilde{F}(x,y) = \mathcal{E}F_P(x,y) = \alpha^*(x,y)$ and the Bayes risk can be computed, being equal to

$$\int_{\Re^2} \{\mathcal{E}[F_P(x,y)]^2 - [\mathcal{E}F_P(x,y)]^2\} dR(x,y).$$

We need to find $\mathcal{E}[F_P(x,y)]^2$ and, applying Theorem 2.2.3, we see that

$$\mathcal{E}[F_P(x,y)]^2 = \frac{\alpha^*(x,y)}{\beta+n+1} + \frac{(\beta+n)(\alpha^*(x,y))^2}{\beta+n+1}.$$

It follows that the Bayes risk is given by

$$1/(\beta+n+1)\int_{\Re^2} \alpha^*(x,y)[1-\alpha^*(x,y)] dR(x,y).$$

The concave function $\alpha^*(x,y)[1-\alpha^*(x,y)]$ is pointwise minimised by taking $\hat{F}$ equal either to $\hat{W}$, if $\hat{W}(x,y)+\hat{M}(x,y) < 1$, or $\hat{M}$, otherwise. Therefore, $\hat{F}$ is a distribution function, which concentrates mass on the line $\hat{W}(x,y)+\hat{M}(x,y) = 1$. The need to have $\hat{F}$ right continuous justifies the choice of $\hat{F}(x,y) = \hat{M}(x,y)$ when $\hat{W}(x,y) + \hat{M}(x,y) = 1$.

$\hat{F}$ is a solution also in $\Pi(\hat{W}, \hat{M})$, because it is obtained by updating, after a sample, a distribution function in $\Gamma(F, G)$:

$$F(x,y) = \begin{cases} W(x,y) & if\ \hat{W}(x,y) + \hat{M}(x,y) < 1 \\ M(x,y) & if\ \hat{W}(x,y) + \hat{M}(x,y) \geq 1. \end{cases}$$

4.3 Nonparametric statistical decisions

In this section, we will consider some criteria to choose an action a in some space, namely a distribution function in the space of all distributions F, when a loss function $L(F, a)$ is specified.

4.3.1 Minimum expected loss criterion

Given a sample of size n from P, we look for the action leading to the lowest expected loss. Since the prior α is specified as a member of the prior Fréchet class $\Gamma(F, G)$, we have to look also for the posterior $\alpha^* \in \Pi(\hat{W}, \hat{M})$ which minimises the expected loss, under a squared loss function.

In estimating a distribution function on $(\Re^2, \mathcal{A})$, let the space $\mathcal{F}$ of all distributions $\tilde{F}$ on $\Re^2$ be the action space and let the loss function be (as in Ferguson, [30])

$$L(F_P, \tilde{F}) = \int_{\Re^2} \{F_P(x, y) - \tilde{F}(x, y)\}^2 dR(x, y),$$

where the weight function R is a given finite measure on $(\Re^2, \mathcal{A})$ and F_P is the random distribution function corresponding to P.

Theorem 4.3.1

$$\mathcal{E}L(F_P, \hat{F}) = \inf_{\alpha^* \in \Pi(\hat{W}, \hat{M})} \inf_{\tilde{F} \in \mathcal{F}} \mathcal{E}L(F_P, \tilde{F}).$$

is obtained by taking

$$\hat{F}(x, y) = \begin{cases} \hat{W}(x, y) & if\ \hat{W}(x, y) + \hat{M}(x, y) < 1 \\ \hat{M}(x, y) & if\ \hat{W}(x, y) + \hat{M}(x, y) \geq 1 \end{cases}$$

$$\begin{aligned}\alpha(x_i,y_i)+\overline{\alpha}(x_i,y_i) &= 2\{1-e^{-x_i}-e^{-y_i}+e^{-(x_i+y_i+\theta x_i y_i)}\}+\\ &+1-\{1-e^{-x_i}\}-\{1-e^{-y_i}\}\\ &=1-e^{-x_i}-e^{-y_i}+2e^{-(x_i+y_i+\theta x_i y_i)}.\end{aligned}$$

As a function of θ, each summand decreases as $e^{-\theta x_i y_i}$ does, so that it achieves its maximum and minimum at, respectively, $\theta = 0$ and $\theta = 1$.

Therefore, the range of the estimator $\hat{\Delta}_B$ equals

$$\frac{\beta+n}{\beta+n+1}\left\{\frac{p_n^2 e^{-2}E_1(2)}{2}+\frac{4p_n(1-p_n)\sum_{i=1}^{n}e^{-(x_i+y_i)}(1-e^{-x_i y_i})}{n}\right\}.$$

□

4.2.3 Positive dependence

Another problem dealt with by Dalal and Phadia [21] was the test of positive dependence versus nonpositive dependence when a bivariate distribution function $H(x,y)$ is given with marginal distributions $F(x)$ and $G(y)$, as discussed in Section 2.3. By using Theorem 2.3.2, we now find upper and lower bounds on

$$\Delta_n(\mathbf{X},\mathbf{Y}) = \int[\frac{\beta+n}{\beta+n+1}H^*(x',y')-\frac{\beta}{\beta+n+1}F^*(x')G^*(y')]dR(x',y').$$

which determines the Bayes rule $\delta(x,y) = I_{\Delta_n(X,Y)\geq 0}$.

Theorem 4.2.4 $\Delta_n(\mathbf{X},\mathbf{Y})$ *is maximised by* $H^* = \hat{M}$ *and minimised by* $H^* = \hat{W}$.

Proof. The only term in $\Delta_n(\mathbf{X},\mathbf{Y})$ changing as H^* varies in $\Pi(\hat{W},\hat{M})$ is $\int H^*(x',y')dR(x',y')$. □

$$\Delta_{(\alpha,\hat{\alpha})} = 1/n \sum_{i=1}^{n} [\alpha(x_i, y_i) + \overline{\alpha}(x_i, y_i)], \tag{4.2.2}$$

where

$$\overline{\alpha}(x_i, y_i) = 1 - \alpha(x_i, \infty) - \alpha(\infty, y_i) + \alpha(x_i, y_i) \text{ and } E_1(x) = \int_1^\infty (e^{-xt}/t)dt.$$

Theorem 4.2.3 *The upper and lower bounds on $\hat{\Delta}_B$, as α varies in $\mathcal{G}$, are obtained for $\theta = 0$ and $\theta = 1$, respectively.*

PROOF. Because of the expression of $\hat{\Delta}_B$, given by equation (2.3.1), we have to look at the behaviour of the quantities Δ_α and $\Delta_{(\alpha,\hat{\alpha})}$ and prove that both of them are decreasing as θ increases.

From (4.2.1) it follows that Δ_α increases as $e^{-2/\theta} E_1(2/\theta) = \int_1^\infty \frac{e^{-2(1+t)/\theta}}{t} dt$ decreases, that is for $\theta \downarrow 0$, since the function $e^{-2(1+t)/\theta}/t$ is positive and has positive derivative, $\frac{2(1+t)e^{-2(1+t)\theta}}{\theta^2 t}$, with respect to θ, so that

$$\int_1^\infty (e^{-2(1+t)/\theta_1}/t)dt \le \int_1^\infty (e^{-2(1+t)/\theta_2}/t)dt \Leftrightarrow \theta_1 \ge \theta_2.$$

Δ_α at $\theta = 0$ can be computed by using the Lebesgue's dominated convergence theorem (Theorem A.1.3), because, for any $\theta \in [0, 1]$,

$$f_\theta(t) = e^{-2(1+t)/\theta}/t \le f_1(\theta) = e^{-2(1+t)}/t,$$

and f_1 is integrable with respect to Lebesgue measure. Besides $f_\theta(t) \to 0$ almost everywhere so that $\int f_\theta(t)dt \to 0$ and $\Delta_\alpha = 1/2$.

By a simple substitution in (4.2.1), it follows that

$$\Delta_\alpha = 1/2 - 1/2e^{-2}E_1(2) \approx .4966910178$$

at $\theta = 1$.

Each summand in (4.2.2), is such that

and

$$\Delta_{\alpha(x_i,y_i)} = P_\alpha\{(X - x_i)(Y - y_i) > 0\} + 1/2 P_\alpha\{(X - x_i)(Y - y_i) = 0\}$$

are such that their second term in the right hand side (r.h.s.) is constant w.r.t. H, depending only on F and G.

The first term of the r.h.s. of Δ_α is obviously maximised by M, which concentrates the mass on a nondecreasing line, and minimised by W, which concentrates the mass on a nonincreasing line. The first term of the r.h.s. of $\Delta_{\alpha(x_i,y_i)}$ equals $P_\alpha\{(X > x_i, Y > y_i)\} + P_\alpha\{(X < x_i, Y < y_i)\}$, i.e. $1 - F(x_i) - G(y_i) + H(x_i, y_i) + H(x_i^-, y_i^-)$, which is maximised by M and minimised by W. □

As a consequence of Theorem 4.2.2, it is possible to find bounds also on Kendall's coefficient, given by $\tau = 2\Delta - 1$.

Upper and lower bounds are computed also for an example presented in Dalal and Phadia [21]. We are considering one of the parametric subclass of the Fréchet class presented in literature (see, e.g., Genest and MacKay [41]).

Example 4.2.1 Consider the class of all α's having a Gumbel bivariate exponential distribution

$$\alpha_\theta(x, y) = 1 - e^{-x} - e^{-y} + e^{-(x+y+\theta xy)},\ 0 \le \theta \le 1.$$

As θ ranges over $[0, 1]$, the α_θ's form a subclass $\mathcal{G}$ of the Fréchet class $\Gamma(F, G)$, where both F and G are exponentially distributed with mean 1. As proved by Dalal and Phadia [21], the Bayes estimator $\hat{\Delta}_B = \mathcal{E}\Delta_H$ is given by equation (2.3.1), with

$$\Delta_\alpha = 1/2 - 1/2 e^{-2/\theta} E_1(2/\theta), \tag{4.2.1}$$

$\sigma = \int[\alpha(x,y) - F(x)G(y)]dxdy$, so that it is maximised by M and minimised by W. $\square$

4.2.2 Concordance

As discussed in Section 2.3, it is worth considering the concordance between pairs of independent observations from the same distribution function $H(x,y)$. When H is chosen according to a Dirichlet process $\mathcal{D}(\eta)$, then the Bayes estimator of the concordance coefficient Δ_H under the squared error loss function is given by Theorem 2.3.1. When α is chosen in the Fréchet class $\Gamma(F,G)$, then it is worth finding upper and lower bounds on the Bayes estimator, to assess, for example, that there is strong concordance among observations if the lower bound exceeds 0.5. Upper and lower bounds are achieved by considering the Fréchet upper and lower bounds, respectively, which give the most concordant probability measure and the most discordant one, as discussed in Chapter 3.

Theorem 4.2.2 *Let $(X_1,Y_1),\dots,(X_n,Y_n)$ be a random sample from a distribution function $H(x,y)$. If H is chosen according to a Dirichlet process $\mathcal{D}(\eta)$, $\eta = \beta\alpha$, $\alpha \in \Gamma(F,G)$, with $P_\alpha\{X = X', Y = Y'\} = 0$ and $P_\alpha\{X = x_i, Y = y_i\} = 0$ for all $(x_i,y_i) \in \Re^2$, then upper and lower bounds on the Bayes estimator of Δ_H, under the squared error loss function, are given by taking $H = M$ and $H = W$, respectively.*

Proof. It suffices to prove that all the quantities specified in Theorem 2.3.1, depending on H, are maximised by taking $H = M$ and minimised by $H = W$. In particular,

$$\Delta_\alpha = P_\alpha\{(X - X')(Y - Y') > 0\} + 1/2 P_\alpha\{(X - X')(Y - Y') = 0\}$$

□

From Theorem 4.1.8, it follows that, whereas the measures $\alpha^* \in \Psi(\hat{W}, \hat{M})$ converge, with probability one, to the true distribution $\tilde{H}$, the measures in the class $\Gamma(F^*, G^*)$ do not, in general, i.e. unless $\min\{\tilde{F}(x), \tilde{G}(y)\} = \max\{\tilde{F}(x) + \tilde{G}(y) - 1, 0\}$. It happens when the measure $\tilde{H}$ is concentrated on lines $x = x_0$ or $y = y_0$ (maybe just in one point) so that at any point (x, y) at least one among $\tilde{F}(x)$ and $\tilde{G}(y)$ equals either 0 or 1.

4.2 Dependence between random variables

In this section we study relations between random variables X and Y when their joint distribution is chosen by a Dirichlet process.

4.2.1 Covariance

Given a sample of size n, Ferguson [30], pag. 227, found out that the Bayes estimator $\hat{\sigma}$ of the covariance of a distribution under squared loss function is given by

$$\hat{\sigma} = \frac{\beta + n}{\beta + n + 1}\{p\sigma + (1-p)s + p(1-p)(\mu_1 - \overline{X})(\mu_2 - \overline{Y})\},$$

where $\mu_1 = \int x dF(x), \mu_2 = \int y dG(y), \sigma = \int xy d\alpha(x, y) - \mu_1\mu_2$, $p = \beta/(\beta + n)$ and s is the sample covariance $\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})/n$.

Theorem 4.2.1 *Upper and lower bounds on $\hat{\sigma}$ are obtained by considering the upper and lower Fréchet bounds.*

Proof. We look for bounds only on σ, the covariance of the distribution $\mathcal{E}P = \alpha$, since it is the unique nonconstant term, as α varies in $\Gamma(F, G)$. As mentioned in Dall'Aglio, Kotz and Salinetti [25], p. 7, Hoeffding proved that with probability one, because the Glivenko-Cantelli theorem (Theorem A.1.1) implies that

$$\lim_{n\to+\infty} \sup_{(x,y)} \left|\frac{\sum_{i=1}^n \delta_{Z_i}(x,y)}{n} - \tilde{H}(x,y)\right| = 0$$

with probability one.

In a similar way, it can be proved that $\lim_{n\to+\infty} \hat{W}(x,y) = \tilde{H}(x,y)$.

The Fréchet upper bound $M^*(x,y)$ is such that, for every $(x,y) \in \Re^2$, $\lim_{n\to+\infty} |M^*(x,y) - \min\{\tilde{F}(x), \tilde{G}(y)\}|$ equals

$$\lim_{n\to+\infty} \left|\min\left\{\frac{\beta F(x) + \sum_{i=1}^n \delta_{X_i}(x)}{\beta+n}, \frac{\beta G(y) + \sum_{i=1}^n \delta_{Y_i}(y)}{\beta+n}\right\} - \min\{\tilde{F}(x), \tilde{G}(y)\}\right|$$

$$\leq \lim_{n\to+\infty} \max\left\{\left|\frac{\beta F(x) + \sum_{i=1}^n \delta_{X_i}(x)}{\beta+n} - \tilde{F}(x)\right|, \left|\frac{\beta G(y) + \sum_{i=1}^n \delta_{Y_i}(y)}{\beta+n} - \tilde{G}(y)\right|\right\}$$

$$\leq \lim_{n\to+\infty} \left|\frac{\beta F(x) + \sum_{i=1}^n \delta_{X_i}(x)}{\beta+n} - \tilde{F}(x)\right| + \lim_{n\to+\infty} \left|\frac{\beta G(y) + \sum_{i=1}^n \delta_{Y_i}(y)}{\beta+n} - \tilde{G}(y)\right|$$

$$\leq \lim_{n\to+\infty} \left|\frac{\sum_{i=1}^n \delta_{X_i}(x)}{n} - \tilde{F}(x)\right| + \lim_{n\to+\infty} \left|\frac{\sum_{i=1}^n \delta_{Y_i}(y)}{n} - \tilde{G}(y)\right| = 0$$

with probability one, because of the Glivenko-Cantelli Theorem, as in the proof of the previous result. The first inequality is proved by considering all the possible combinations of minimising functions at any $(x,y) \in \Re^2$, and then majorising in order to get the above majorisation. The last inequality is proved by using the triangle inequality and observing that

$$\lim_{n\to+\infty} \frac{\beta F(x)}{\beta+n} = \lim_{n\to+\infty} \frac{\beta G(y)}{\beta+n} = 0.$$

In a similar way, it can be proved that

$$\lim_{n\to+\infty} W^*(x,y) = \max\{\tilde{F}(x) + \tilde{G}(y) - 1, 0\}.$$

Theorem 4.1.7 *For every* $(x,y) \in \Re^2$,

$$\lim_{\beta\to 0} \hat{M}(x,y) = \lim_{\beta\to 0} \hat{W}(x,y) = \frac{\sum_{i=1}^{n} \delta_{Z_i}(x,y)}{n},$$

$$\lim_{\beta\to 0} W^*(x,y) = \max\{\frac{\sum_{i=1}^{n} \delta_{X_i}(x)}{n} + \frac{\sum_{i=1}^{n} \delta_{Y_i}(y)}{n} - 1, 0\}$$

and

$$\lim_{\beta\to 0} M^*(x,y) = \min\{\frac{\sum_{i=1}^{n} \delta_{X_i}(x)}{n}, \frac{\sum_{i=1}^{n} \delta_{Y_i}(y)}{n}\}.$$

Different results are obtained, in general, when β is kept fixed whereas the sample size n goes to infinity. Consider a sequence of observations $Z_1 = (X_1, Y_1), Z_2 = (X_2, Y_2), \ldots$ assumed to be independent and identically distributed (i.i.d.) with joint distribution $\tilde{H}(x,y)$ and marginals $\tilde{F}(x)$ and $\tilde{G}(y)$.

Theorem 4.1.8 *For every* $(x,y) \in \Re^2$, *a.s.,*

$$\lim_{n\to+\infty} \hat{M}(x,y) = \lim_{n\to+\infty} \hat{W}(x,y) = \tilde{H}(x,y),$$

$$\lim_{n\to+\infty} M^*(x,y) = \min\{\tilde{F}(x), \tilde{G}(y)\},$$

and

$$\lim_{n\to+\infty} W^*(x,y) = \max\{\tilde{F}(x) + \tilde{G}(y) - 1, 0\}.$$

Proof. The upper bound $\hat{M}(x,y)$ is such that, for every $(x,y) \in \Re^2$, $\lim_{n\to+\infty} |\hat{M}(x,y) - \tilde{H}(x,y)| =$

$$\begin{aligned}
&= \lim_{n\to+\infty} |\frac{\beta}{\beta+n} \min\{F(x), G(y)\} + \frac{\sum_{i=1}^{n} \delta_{Z_i}(x,y)}{\beta+n} - \tilde{H}(x,y)| \\
&\leq \lim_{n\to+\infty} \frac{\beta}{\beta+n} \min\{F(x), G(y)\} + \lim_{n\to+\infty} |\frac{\sum_{i=1}^{n} \delta_{Z_i}(x,y)}{\beta+n} - \tilde{H}(x,y)| \\
&= \lim_{n\to+\infty} |\frac{\sum_{i=1}^{n} \delta_{Z_i}(x,y)}{n} - \tilde{H}(x,y)| = 0
\end{aligned}$$

PROOF. The result follows from Theorem 4.1.6 and the triangle inequality, because

$$|W^*(x,y) - \hat{W}(x,y)| \leq |W^*(x,y) - W(x,y)| + |W(x,y) - \hat{W}(x,y)| \to 0,$$

as $\beta \to +\infty$. The result about M^* and $\hat{M}$ is proved similarly. □

Corollary 4.1.5 *For a given sample size n and any $\delta > 0$, then $\beta \geq n(1-\delta)/\delta$ implies that $\{M^*(x,y) - \hat{M}(x,y)\} \leq \delta$ and $\{\hat{W}(x,y) - W^*(x,y)\} \leq \delta$ for all $(x,y) \in \Re^2$.*

PROOF. We prove the result when $F(x) \leq G(y)$; the other case may be treated in a similar way.

Two different situations are to be faced, depending on the possible values M^* can take. Suppose, first, that $M^*(x,y) = \dfrac{\beta F(x) + \sum_{i=1}^n \delta_{X_i}(x)}{\beta + n}$, so that

$$\begin{aligned} M^*(x,y) - \hat{M}(x,y) &= \frac{\beta F(x) + \sum_{i=1}^n \delta_{X_i}(x)}{\beta + n} - \frac{\beta F(x) + \sum_{i=1}^n \delta_{Z_i}(x,y)}{\beta + n} \\ &\leq \frac{n}{\beta + n}. \end{aligned}$$

In the other case, it follows that

$$\begin{aligned} M^*(x,y) - \hat{M}(x,y) &= \frac{\beta G(y) + \sum_{i=1}^n \delta_{Y_i}(y)}{\beta + n} - \frac{\beta F(x) + \sum_{i=1}^n \delta_{Z_i}(x,y)}{\beta + n} \\ &\leq \frac{\beta F(x) + \sum_{i=1}^n \delta_{X_i}(x)}{\beta + n} - \frac{\beta F(x) + \sum_{i=1}^n \delta_{Z_i}(x,y)}{\beta + n} \\ &\leq \frac{n}{\beta + n}. \end{aligned}$$

The last term in the inequality is less than δ when $\beta \geq n(1-\delta)/\delta$.

The proof about the lower bounds is similar. □

As the "strength of belief" β goes to 0, then the bounds converge to the empirical distributions, as stated in the next Theorem, whose simple proof is omitted.

We study now the relation between $\Psi(\hat{W},\hat{M})$ and $\Gamma(F^*,G^*)$ when the sample size increases or the "strength of belief" β goes either to zero or to infinity.

Theorem 4.1.6 *For every* $(x,y)\in\Re^2$*,*

$$\lim_{\beta\to+\infty} W^*(x,y) = \lim_{\beta\to+\infty} \hat{W}(x,y) = W(x,y)$$

and

$$\lim_{\beta\to+\infty} M^*(x,y) = \lim_{\beta\to+\infty} \hat{M}(x,y) = M(x,y).$$

Proof. We prove the result only for the posterior upper Fréchet bound M^* since the proofs for W^*, $\hat{W}$ and $\hat{M}$ are similar.

Consider a point (x,y), then, as $\beta\to+\infty$, it can be easily proved that $F(x)<G(y)$ implies $M^*(x,y)=F^*(x)$. The empirical distribution is thus irrelevant, as expected, in determining the asymptotic behaviour. For every $(x,y)\in\Re^2$, it follows that

$$\begin{aligned}|M(x,y)-M^*(x,y)| &= |F(x)-\frac{\beta F(x)+\sum_{i=1}^n \delta_{X_i}(x)}{\beta+n}| \\ &= |\frac{nF(x)-\sum_{i=1}^n \delta_{X_i}(x)}{\beta+n}| \\ &\le \frac{n}{\beta+n}\end{aligned}$$

and the last term goes to 0 as β goes to infinity.

The cases $F(x)>G(y)$, implying $M^*(x,y)=G^*(x,y)$, and $F(x)=G(y)$ are treated similarly. □

Corollary 4.1.4 *For every* $(x,y)\in\Re^2$*, it follows that*

$$\lim_{\beta\to+\infty}\{W^*(x,y)-\hat{W}(x,y)\}=0 \text{ and } \lim_{\beta\to+\infty}\{M^*(x,y)-\hat{M}(x,y)\}=0.$$

and

$$\hat{M}(x,y) = \frac{\beta}{\beta+n} M(x,y) + \frac{\sum_1^n \delta_{Z_i}(x,y)}{\beta+n}.$$

□

An important measure of the distance between upper and lower bounds is given by the following result, which holds under very mild conditions (e.g. F and G continuous and strictly increasing). Such a result could be proved in more general cases, with ad hoc changes, as in the case considered in Chapter 6.

Theorem 4.1.5 *Suppose that there exists* $(\hat{x},\hat{y}) \in \Re^2$ *such that* $F(\hat{x}) = G(\hat{y}) = 1/2$*, then it follows that*

$$\sup_{(x,y)\in\Re^2} \{\hat{M}(x,y) - \hat{W}(x,y)\} = \frac{\beta}{2(\beta+n)}.$$

Proof. Because of Theorem 4.1.4, we must look at the prior Fréchet bounds only. We consider the following subsets of $\Re^2$, whose union gives $\Re^2$:

$$T_1 = \{(x,y) : W(x,y) \leq 0, M(x,y) = F(x)\},$$

$$T_2 = \{(x,y) : W(x,y) \geq 0, M(x,y) = F(x)\},$$

$$T_3 = \{(x,y) : W(x,y) \leq 0, M(x,y) = G(y)\},$$

$$T_4 = \{(x,y) : W(x,y) \geq 0, M(x,y) = G(y)\}.$$

These subsets have a unique common point: $(\hat{x},\hat{y})$ such that $F(\hat{x}) = G(\hat{y}) = 1/2$. It can be proved that, in each $T_i, i = 1,\ldots,4$, $\sup_{(x,y)\in\Re^2} \{M(x,y) - W(x,y)\}$ is achieved at $(\hat{x},\hat{y})$. Since $M(\hat{x},\hat{y}) = 1/2$ and $W(\hat{x},\hat{y}) = 0$, the result is proved. □

4.1.2 Expectations

When computing expectations of functions of either X or Y in $\Gamma(F^*, G^*)$, the same result is obtained, regardless of the joint distribution α. In fact, as an example, note that

$$EX = \int x dF^*(x) \text{ and } VarX = \int x^2 dF^*(x) - \{EX\}^2,$$

whereas the covariance (and the correlation, too) is within the upper and lower bounds determined by the Fréchet bounds (or by $\hat{W}$ and $\hat{M}$ in $\Psi(\hat{W}, \hat{M})$ and $\Pi(\hat{W}, \hat{M})$).

It can be easily proved a result similar to Theorem 3.1.2 about the minimisation of $E|X-Y|^\nu, \nu \geq 1$, being the sample uninfluential on the minimisation.

Theorem 4.1.3 *$E|X-Y|^\nu, \nu \geq 1$, is minimised over $\Psi(\hat{W}, \hat{M})$ by considering in it the measures corresponding to the minimising ones on $\Gamma(F, G)$.*

4.1.3 Relations between bounds

The distance between upper and lower bounds decreases as the sample size increases and such decrement is not affected by the observed Z_i's, as proved by the following result.

Theorem 4.1.4

$$\hat{M}(x,y) - \hat{W}(x,y) = \frac{\beta}{\beta+n}\{M(x,y) - W(x,y)\}, \forall (x,y) \in \Re^2.$$

PROOF. It follows immediately from observing that

$$\hat{W}(x,y) = \frac{\beta}{\beta+n} W(x,y) + \frac{\sum_1^n \delta_{Z_i}(x,y)}{\beta+n}$$

the following example.

Example 4.1.2 Let P be a Dirichlet process on $([0,1] \times [0,1], \mathcal{A})$ with parameter η such that $\beta = 1$ and $\alpha \in \Gamma(F, G)$, where F and G are uniform distribution functions. Given a sample $Z_1 = (x_1, y_1)$, then $\alpha^* = (\alpha + \delta_{Z_1})/2$. The marginal distributions of α^* are mixtures, with equal weights $1/2$, of a uniform distribution and a point mass at x_1 and y_1, respectively. Consider the measure α_0 which spreads uniformly its mass all over $\{[0,1] \times y_1\}$ and $\{x_1 \times [0,1]\}$: it has the same marginal distributions as α^* but it does not come from $\Gamma(F, G)$. Therefore $\alpha_0 \in \Gamma(F^*, G^*)$, but $\alpha_0 \notin \Pi(\hat{W}, \hat{M})$.

Besides, the distribution function α_0 is not within the band determined by $\hat{W}$ and $\hat{M}$, unless $x_1 = y_1 = 1/2$. In fact, suppose that $x_1 + y_1 > 1$ and consider $\hat{x}$ and $\hat{y}$ such that $\hat{x} < x_1$, $\hat{y} < y_1$ and $\hat{x} + \hat{y} > 1$, therefore $\alpha_0(\hat{x}, \hat{y}) = 0 < (\hat{x} + \hat{y} - 1)/2 = \hat{W}(\hat{x}, \hat{y})$. When $x_1 + y_1 < 1$, then it follows that $\alpha_0(\hat{x}, \hat{y}) = (\hat{x} + \hat{y})/2 < 1/2 = \hat{W}(\hat{x}, \hat{y})$, when $\hat{x} > x_1$, $\hat{y} > y_1$ and $\hat{x} + \hat{y} < 1$. Therefore, α_0 might be within the band only if $x_1 + y_1 = 1$.

Looking at $\hat{M}$ and M^*, consider (x_1, y_1), with $x_1 < y_1$, and $\hat{x}$ and $\hat{y}$ such that $\hat{x} > x_1$, $\hat{y} < y_1$ and $\hat{x} < \hat{y}$, therefore $\alpha_0(\hat{x}, \hat{y}) = \hat{y}/2 > \hat{x}/2 = \hat{M}(\hat{x}, \hat{y})$. When $x_1 > y_1$, then it follows that $\alpha_0(\hat{x}, \hat{y}) = \hat{x}/2 < \hat{y}/2 = \hat{M}(\hat{x}, \hat{y})$, when $\hat{x} < x_1$, $\hat{y} > y_1$ and $\hat{x} > \hat{y}$. Therefore, α_0 might be within the band only if $x_1 = y_1$.

The two previous conditions can be combined into $x_1 = y_1 = 1/2$ and then it can be easily proved that $\Psi(\hat{W}, \hat{M})$ and $\Gamma(F^*, G^*)$ coincide; should the sample be different, then $\Psi(\hat{W}, \hat{M})$ would be strictly contained in $\Gamma(F^*, G^*)$.

Suppose that τ and ν intersect only in one point $(\hat{x}, \hat{y})$. Looking at the conditions defining T_1, T_2, T_3 and T_4, it follows that $(\hat{x}, \hat{y})$ is such that $F(\hat{x}^-) < 1/2, F(\hat{x}) \geq 1/2, G(\hat{y}^-) < 1/2$ and $G(\hat{y}) \geq 1/2$. It can be easily seen that both upper and lower bounds coincide if all the observations equal $(\hat{x}, \hat{y})$.

Suppose that τ and ν intersect on a vertical segment in the plane (X, Y). Looking at the conditions defining T_1, T_2, T_3 and T_4, it follows that all the points (x, y) on the segment are such that $F(x^-) < 1/2, F(x) \geq 1/2$ and the bounds coincide if all the observations are in the segment.

Suppose that τ and ν intersect on a horizontal segment in the plane (X, Y). Looking at the conditions defining T_1, T_2, T_3 and T_4, it follows that all the points (x, y) on the segment are such that $G(y^-) < 1/2, G(y) \geq 1/2$ and the bounds coincide if all the observations are in the segment.

Because of the definition of $\Psi(\hat{W}, \hat{M})$, it coincides with $\Gamma(F^*, G^*)$ when $\hat{W}(x, y) = W^*(x, y)$ and $\hat{M}(x, y) = M^*(x, y)$ for all $(x, y) \in \Re^2$. $\Box$

The next Corollary is an immediate, important consequence of Theorem 4.1.2.

Corollary 4.1.2 *If the distribution functions F and G are continuous and strictly increasing, then $\Psi(\hat{W}, \hat{M})$ and $\Gamma(F^*, G^*)$ coincide if and only if all the observations equal $(\hat{x}, \hat{y})$ such that $F(\hat{x}) = G(\hat{y}) = 1/2$.*

Consider a sequence of observations $Z_1 = (X_1, Y_1), Z_2 = (X_2, Y_2), \ldots$ assumed to be independent and identically distributed (i.i.d.) with joint distribution $\tilde{H}(x, y)$ and continuous, increasing marginals $\tilde{F}(x)$ and $\tilde{G}(y)$. It can be easily proved the following

Corollary 4.1.3 $\Psi(\hat{W}, \hat{M})$ *is strictly contained in* $\Gamma(F^*, G^*)$ *a.s..*

The class $\Psi(\hat{W}, \hat{M})$ may be strictly contained in $\Gamma(F^*, G^*)$, as shown by necessary: for any (x, y) such that $F(x) \leq G(y)$, there should be no sample $Z_i = (x_i, y_i)$ with $x_i \leq x$ and $y_i > y$, and, for any (x, y) such that $F(x) \geq G(y)$, there should be no sample $Z_i = (x_i, y_i)$ with $x_i > x$ and $y_i \leq y$.

Suppose that there is an observation (x_i, y_i) in the interior of T_1 (T_2 is treated similarly), thus there exists $\varepsilon > 0$ such that $(x_i, y_i - \varepsilon)$ is inside T_1 but it does not satisfy the above necessary conditions, so that $M^*(x_i, y_i - \varepsilon) > \hat{M}(x_i, y_i - \varepsilon)$. Therefore, we have to look only at $(x_i, y_i) \in \tau$.

We are now going to look for conditions on F, G and the sample leading to $W^*(x, y) = \hat{W}(x, y)$ for all $(x, y) \in \Re^2$. We can prove that two conditions are necessary: for any (x, y) such that $F(x) + G(y) \geq 1$, there should be no sample $Z_i = (x_i, y_i)$ with $x_i > x$ and $y_i > y$, and, for any (x, y) such that $F(x) + G(y) \leq 1$, there should be no sample $Z_i = (x_i, y_i)$ with $x_i \leq x$ and $y_i \leq y$.

Suppose that there is an observation (x_i, y_i) in the interior of T_3, thus $W^*(x_i, y_i) < \hat{W}(x_i, y_i)$. Let (x_i, y_i) be in the interior of T_4, thus there exists $\varepsilon > 0$ such that $(x_i - \varepsilon, y_i - \varepsilon)$ is inside T_4 but it does not satisfy the above necessary conditions, so that $W^*(x_i - \varepsilon, y_i - \varepsilon) < \hat{W}(x_i - \varepsilon, y_i - \varepsilon)$. Therefore, we have to look only at $(x_i, y_i) \in \nu$.

Let $x_0 < x$ and $y_0 < y$ be such that both (x_0, y_0) and (x, y) are in $\tau \cap \nu$, so that it can be proved that $Q = [x_0, x] \times [y_0, y]$ is in $\tau \cap \nu$ and the above necessary conditions are not satisfied in Q. Therefore, we can have neither observations in any rectangle common to τ and ν nor τ and ν intersecting at more than one isolated point, because they contain also the joining segment. It follows that the bounds might coincide only where τ and ν intersect, either in an isolated point or in a segment.

Let τ be the line separating

$$T_1 = \{(x,y) : F(x) < G(y)\} \text{ from } T_2 = \{(x,y) : F(x) > G(y)\},$$

and ν the line separating

$$T_3 = \{(x,y) : F(x) + G(y) < 1\} \text{ from } T_4 = \{(x,y) : F(x) + G(y) > 1\}.$$

It can be seen that $\tau \equiv \{(x,y) : F(x) = G(y)\}$ if and only if F and G are either both continuous or with jumps only at (x,y) such that $F(x^-) = G(y^-)$ and $F(x) = G(y)$.

On the other hand, $\nu \equiv \{(x,y) : F(x) + G(y) = 1\}$ if and only if F and G are both continuous; in this case jumps are not allowed since the right continuity of the distribution functions implies $G(y^+) = G(y)$ so that $F(x^-) + G(y^+) = F(x) + G(y)$ if and only if $F(x) = F(x^-)$.

Theorem 4.1.2 *The class $\Psi(\hat{W}, \hat{M})$ is contained in $\Gamma(F^*, G^*)$. The inclusion is strict unless the observations are in the intersection of τ and ν, but only if it is either an isolated point or a segment.*

Proof. We prove that $\Psi(\hat{W}, \hat{M})$ is contained in $\Gamma(F^*, G^*)$ by showing that any $\alpha^* \in \Psi(\hat{W}, \hat{M})$ has marginals F^* and G^*.

$$\alpha^*(x,\infty) = \frac{\beta\alpha(x,\infty) + \sum_1^n \delta_{Z_i}(x,\infty)}{\beta + n} = \frac{\beta F(x) + \sum_1^n \delta_{X_i}(x)}{\beta + n} = F^*(x).$$

Similarly, $\alpha^*(\infty, y) = G^*(y)$.

In the next, we consider only distinct values of the observations, being the proof unaffected by coincident ones.

We are now going to look for conditions on F, G and the sample leading to $M^*(x,y) = \hat{M}(x,y)$ for all $(x,y) \in \Re^2$. We can prove that two conditions are difference w.r.t. what we have proved. □

Consider a sequence of observations $Z_1 = (X_1, Y_1), Z_2 = (X_2, Y_2), \ldots$ assumed to be independent and identically distributed (i.i.d.) with joint distribution $\tilde{H}(x, y)$ and continuous, increasing marginals $\tilde{F}(x)$ and $\tilde{G}(y)$. It can be easily proved the following

Corollary 4.1.1 $\Pi(\hat{W}, \hat{M})$ *is strictly contained in* $\Psi(\hat{W}, \hat{M})$ *a.s..*

Example 4.1.1 Consider $\beta = n = 1$ and the sample $Z_1 = (z, z)$. Let F and G be uniform distributions on the unit square, so that

$$\hat{W}(x, y) = \frac{\max\{x + y - 1, 0\} + \delta_{Z_1}(x, y)}{2},$$

and

$$\hat{M}(x, y) = \frac{\min\{x, y\} + \delta_{Z_1}(x, y)}{2}.$$

Let $S = \{(x, y) \in [0, 1] \times [0, 1] : x \geq z, y \geq z\}$. Consider the measure $\hat{\alpha}$ which coincides with $\hat{W}$ on S and with $\hat{M}$ otherwise. Being such a measure within the band determined by $\hat{W}$ and $\hat{M}$, therefore it is in $\Pi(\hat{W}, \hat{M})$. The measure $\hat{\alpha}$ gives mass $z/2$ to the point (z, z), uniformly spreads mass $z/2$ on $T_1 = \{(t, t) : 0 \leq t < z\}$ and mass $(1 - z)/2$ on both $T_2 = \{(t, z) : z < t \leq 1\}$ and $T_3 = \{(z, t) : z < t \leq 1\}$.

Suppose that $\hat{\alpha}$ comes from a measure α updated after getting the sample Z_1, then it follows that $\alpha(x, y) = \{x + y - 1\})/2$ for $(x, y) \in S$ and $\alpha(x, y) = \min\{x/2, y/2\}$ otherwise. Since $x/2 > (x + y - 1)/2$ in the interior of S, it follows that α is not a distribution function and therefore $\Pi(\hat{W}, \hat{M}) \subset \Psi(\hat{W}, \hat{M})$.

The class $\Psi(\hat{W}, \hat{M})$ is not a Fréchet class, but it is contained (strictly, in general) in one, as shown by the next result.

then checking that the considered measures are in $\Pi(\hat{W}, \hat{M})$ and the proofs are not affected by dealing with the smaller class.

Theorem 4.1.1 *Let F and G be continuous and increasing. The class $\Pi(\hat{W}, \hat{M})$ is contained in $\Psi(\hat{W}, \hat{M})$. The inclusion is strict unless the observations are in (x_i, y_i) such that $W(x_i, y_i) = M(x_i, y_i)$, i.e.* $\min\{F(x_i), G(y_i), 1 - F(x_i), 1 - G(y_i)\} = 0$.

PROOF. From their definitions, it follows that $\Pi(\hat{W}, \hat{M})$ is obviously contained in $\Psi(\hat{W}, \hat{M})$. We look now for conditions under which the two classes coincide, i.e. such that, given any $\alpha^* \in \Psi(\hat{W}, \hat{M})$, it is obtained by updating some $\alpha \in \Pi(\hat{W}, \hat{M})$ after observing the sample.

Therefore, such α should be equal to $\dfrac{(\beta+n)\alpha^* - \sum_{i=1}^n \delta_{Z_i}}{\beta}$ and be bounded by the prior Fréchet bounds. Such α is right continuous, being sum of right continuous functions. Besides, we must prove that $\alpha(x_2, y_2) - \alpha(x_1, y_2) - \alpha(x_2, y_1) + \alpha(x_1, y_1) \geq 0$ for all $x_1 \leq x_2$, $y_1 \leq y_2$. It can be easily proved if no observation falls in the rectangle with vertices $(x_1, y_1), (x_2, y_1), (x_1, y_2)$ and (x_2, y_2), since it reduces to prove the same condition about α^*, which is, of course, satisfied, being α^* a distribution function. We can see that we are left with the case of the observation being in the upper right corner (x_2, y_2). We have to consider the critical case of α^* having a jump at (x_2, y_2), so that we want to find conditions under which $\Delta_\alpha = \alpha(x_2, y_2) - \alpha(x_2^-, y_2^-) \geq 0$: it follows $\Delta_\alpha \geq \dfrac{k}{\beta + n}$, where k is the number of the observations at (x_2, y_2). Therefore, to avoid having α such that $\Delta_\alpha < 0$, it must be $\hat{W}(x_2, y_2) - \hat{M}(x_2^-, y_2^-) \geq \dfrac{k}{\beta + n}$, which implies, being F and G continuous, that $\hat{W}(x_2, y_2) = \hat{M}(x_2, y_2)$, completing the proof. We might have considered $\hat{M}(x_2^-, y_2)$ but the continuity of F and G makes no

In the next, we will use α to denote both a probability measure and a distribution function, being always evident which one we are referring to.

4.1.1 Classes of probability measures

Suppose that $\alpha \in \Gamma(F, G)$, i.e. let α be in the Fréchet class with given marginals F and G. Let $Z_1, \ldots, Z_n$, $Z_i = (X_i, Y_i), i = 1, \ldots, n$, be a sample of size n from P so that, from Theorem 2.2.1, the conditional distribution of P given $Z_1, \ldots, Z_n$ is a Dirichlet process with parameter $\eta^* = (\beta + n)\alpha^*$, where $\alpha^* = (\beta\alpha + \sum_1^n \delta_{Z_i})/(\beta + n)$, $\alpha^*(\Re^2) = 1$. Lower and upper Fréchet bounds in $\Gamma(F, G)$ are updated into

$$\hat{W}(x, y) = \frac{\beta \max\{F(x) + G(y) - 1, 0\} + \sum_1^n \delta_{Z_i}(x, y)}{\beta + n},$$

and

$$\hat{M}(x, y) = \frac{\beta \min\{F(x), G(y)\} + \sum_1^n \delta_{Z_i}(x, y)}{\beta + n}.$$

Let $\Pi(\hat{W}, \hat{M})$ be the class of all distribution functions obtained by updating $\Gamma(F, G)$; let $\Psi(\hat{W}, \hat{M})$ be the class of all distribution functions α such that

$$\hat{W}(x, y) \leq \alpha(x, y) \leq \hat{M}(x, y), \forall (x, y) \in \Re^2,$$

and let $\Gamma(F^*, G^*)$ be the Fréchet class whose marginals are

$$F^*(x) = \frac{\beta F(x) + \sum_1^n \delta_{X_i}(x)}{\beta + n} \text{ and } G^*(y) = \frac{\beta G(y) + \sum_1^n \delta_{Y_i}(y)}{\beta + n}.$$

We will prove that $\Pi(\hat{W}, \hat{M}) \subseteq \Psi(\hat{W}, \hat{M}) \subseteq \Gamma(F^*, G^*)$, finding conditions ensuring that the inclusions are proper. Besides, we will prove some results by looking for measures α in $\Psi(\hat{W}, \hat{M})$, more manageable than $\Pi(\hat{W}, \hat{M})$, and

Chapter 4

Dirichlet process and Fréchet class

Many statistical problems have been faced in literature by means of Dirichlet processes; in this chapter, we are going to deal with some of them (e.g. dependence between random variables, distance between distribution functions, estimation of distribution functions) when the uncertainty on the parameter, a measure on $\Re^2$, can be specified by means of a Fréchet class $\Gamma(F,G)$. Furthermore, we will study the properties of some classes obtained by updating $\Gamma(F,G)$ after observing a sample from the Dirichlet process.

4.1 Updating the class of parameters

We consider Dirichlet processes P whose parameters are proportional to probability measures in a Fréchet class $\Gamma(F,G)$. Given a sample from P, new classes are obtained and their properties are studied, being of interest both by themselves and for their use in the following chapters.

Consider a Dirichlet process P on $(\Re^2,\mathcal{A})$ with parameter η, where

$$\eta(x,y)=\beta\alpha(x,y) \text{ and } \alpha(\Re^2)=1.$$

Theorem 3.1.2 (Dall'Aglio [22]) *For $\nu > 1$, the quantity $E|X-Y|^\nu$ is minimised over $\Gamma(F,G)$ by taking $H(x,y) = M(x,y)$.*

For $\nu = 1$, $E|X-Y|$ is minimised by any distribution function achieving the maximum value, among the distributions in $\Gamma(F,G)$, on the line $x = y$. In particular, such set of distributions contains M as a maximum element, i.e. any other distribution does not exceed $M(x,y)$ at any point (x,y), whereas the minimum element is given by

$$H^*(x,y) = \begin{cases} F(x) - \max\{\inf_{x \le z \le y} [F(z) - G(z)], 0\} & x \le y \\ G(y) - \max\{\inf_{y \le z \le x} [G(z) - F(z)], 0\} & x \ge y. \end{cases}$$

As proved by Bertino [8], H^* assigns the maximum probability, compatible with the specified marginals, to the line $x = y$. For completeness, we present also the following result.

Theorem 3.1.3 (Bertino [9]) *The expectation of the distance $d(X,Y) = f(|X-Y|)$ is minimised, among the distributions in $\Gamma(F,G)$, by M when f is convex and by H^* when f is concave.*

where $W(x,y) = max\{F(x)+G(y)-1,0\}$ *and* $M(x,y) = min\{F(x),G(y)\}$ *are elements of* $\Gamma(F,G)$, *called Fréchet bounds.*

W and M represent joint distributions which are respectively "statistically decreasing" and "statistically increasing", i.e. they are concentrated on a non-increasing (nondecreasing) curve.

The Fréchet class can be extended to n–dimensional distribution functions, when unidimensional or k–dimensional, $k < n$, marginals are specified. As an example, three and four dimensional distribution functions, with unidimensional marginals, have been considered by Feron [34] and Rizzi [52], respectively. Also in this case, Fréchet bounds are found, but the lower bound is no more a proper distribution function, unless very restrictive conditions are satisfied, as presented in Dall'Aglio [23].

As mentioned in Dall'Aglio [24], the distance between two random variables X and Y having given distribution functions F and G, respectively, can be expressed by $d(X,Y) = f(|X-Y|)$, where $f(z)$ is an increasing sub-additive function for $z \geq 0$, such that $f(0) = 0$, as considered by Fréchet [40]. The main results are due to Dall'Aglio [22] and refer to the choice $f(z) = z^{\nu}$. He considered

$$E|X-Y| = \int_{\Re}[F(z)+G(z)-2H(z,z)]dz,$$

and, for $\nu > 1$,

$$\begin{aligned} E|X-Y|^{\nu} = \ & \nu(\nu-1)\int_{u>v}[G(v)-H(u,v)](u-v)^{\nu-2}dudv+ \\ & \nu(\nu-1)\int_{u<v}[F(u)-H(u,v)](v-u)^{\nu-2}dudv, \end{aligned}$$

where H is a distribution function in $\Gamma(F,G)$.

Chapter 3

Fréchet classes

When considering random vectors, sometimes it is worth studying all possible probability measures having given marginal distributions, i.e. in a Fréchet class. We are going to consider only two-dimensional random variables, but we could consider, as well, n–dimensional random variables. We present now the properties of the Fréchet classes which are going to be used in the subsequent chapters. Most of the results can be found, along with references to their original publication, in the book edited by Dall'Aglio, Kotz and Salinetti [25].

3.1 Definition and main properties

Consider the space $\Re^2$ and the Borel σ–field $\mathcal{A}$ on it. Let F and G be two distributions on $\Re$ and let $\mathcal{B}$ denote the Borel σ–field on the real line.

Definition 3.1.1 *The class of all bivariate distribution functions with given marginals F and G is called the Fréchet class $\Gamma(F, G)$.*

Theorem 3.1.1 (Fréchet [39]) *Given any distribution function H in the class $\Gamma(F, G)$, then it follows that*

$$W(x, y) \leq H(x, y) \leq M(x, y), \forall (x, y) \in \Re^2,$$

Corollary 2.4.1 *Given any weight function R, the action a_0 is chosen if*

$$\sup_{x\in\Re}\left\{\frac{\beta F(x)+m\hat{F}_m(x)}{\beta+m}-\frac{\beta G(x)+n\hat{G}_n(x)}{\beta+n}\right\}<\varepsilon.$$

When $m=n$ and $F=G$, the dependency on α disappears.

Corollary 2.4.2 *Given any weight function R and $n=m$, the action a_0 is chosen if*

$$\sup_{x\in\Re}\{\hat{F}_n(x)-\hat{G}_n(x)\}<\varepsilon\frac{\beta+n}{n}.$$

Finally, we consider $\beta\downarrow 0$ so that we get the one-sided Kolmogorov-Smirnov statistics. The simple proof is omitted.

Corollary 2.4.3 *Given any weight function R, the one-sided Kolmogorov-Smirnov statistics is a limit of the Bayes one, since the action a_0 is chosen if*

$$\sup_{x\in\Re}\{\hat{F}_m(x)-\hat{G}_n(x)\}<\varepsilon.$$

The other one-sided test is treated similarly, as well as the Kolmogorov-Smirnov statistics for the one-sample case, where $\mathcal{E}\tilde{G}=\tilde{G}$.

Finally, it can be shown that the two one-sided test can be combined, and the following result holds.

Corollary 2.4.4 *Given any weight function R, the two-sided Kolmogorov-Smirnov statistics is a limit of the Bayes one, since the action a_0 is chosen if*

$$\sup_{x\in\Re}|\hat{F}_m(x)-\hat{G}_n(x)|<\varepsilon.$$

It should be noticed that $L(\alpha, a_1)$ could be positive even if there exists $x \in \Re$ such that $\tilde{F}(x) - \tilde{G}(x) > \varepsilon$, but it is not a problem since we are looking for sufficient conditions to accept a_0. In fact, we are looking for conditions valid for every R, so that it follows that $L(\alpha, a_1) = 0$ for the weight functions concentrated on the points where $\tilde{F}(x) - \tilde{G}(x) > \varepsilon$. Therefore such loss function, very suitable in computing $\Delta_n(X, Y)$ in the next, reaches its goal of being equal to 0 if the alternative hypothesis is true.

For given pairs of (vectorial) observations $(\mathbf{x}, \mathbf{y})$, we choose the action a_0 if $\Delta_n(X, Y) \leq 0$, where

$$\begin{aligned}
\Delta_n(\mathbf{X}, \mathbf{Y}) &= \mathcal{E}[L(\alpha, a_0) - L(\alpha, a_1)|(\mathbf{X}, \mathbf{Y})] \\
&= \int \mathcal{E}[\tilde{F}(x) - \tilde{G}(x) - \varepsilon|(\mathbf{X}, \mathbf{Y})]dR(x) \\
&= \int \left\{ \frac{\beta F(x) + m\hat{F}_m(x)}{\beta + m} - \frac{\beta G(x) + n\hat{G}_n(x)}{\beta + n} - \varepsilon \right\} dR(x),
\end{aligned}$$

where the expectation is taken w.r.t. the posterior Dirichlet process and $\hat{F}_m$ and $\hat{G}_n$ are the empirical distributions.

Without loss of generality, we may assume $\int dR(x) = 1$ so that we get the following result.

Theorem 2.4.1 *Given the weight function R, the action a_0 is chosen if*

$$\int \left\{ \frac{\beta F(x) + m\hat{F}_m(x)}{\beta + m} - \frac{\beta G(x) + n\hat{G}_n(x)}{\beta + n} \right\} dR(x) < \varepsilon.$$

Theorem 2.4.1 depends on the chosen weight function R. If we want independence on the weight function, as discussed above, then the following result holds.

It is possible to consider the Kolmogorov-Smirnov statistics to compare two populations, by testing the equality of two distribution functions by means of empirical distributions. The two-samples analog of D_n is given by

$$D_{m,n} = \sup_{x \in \Re} |\hat{F}_m(x) - \hat{G}_n(x)|,$$

whereas the analog of D_n^+ and D_n^- are defined similarly.

We now present a nonparametric Bayesian interpretation of the one-sided Kolmogorov-Smirnov statistics, in both one-sample and two-samples cases.

We are interested in comparing the distribution functions $\tilde{F}$ and $\tilde{G}$ of two random variables. Consider a Dirichlet process on $(\Re^2, \mathcal{A})$, with parameter $\eta = \beta\alpha$, with $\beta > 0$ and α a probability measure having marginals F and G (the result depends only on the marginals and not on the joint structure).

We consider the null hypothesis

$$H_0 : \tilde{F}(x) - \tilde{G}(x) \leq \varepsilon \ \forall x \in \Re$$

against the alternative one

$$H_1 : \tilde{F}(x) - \tilde{G}(x) > \varepsilon \text{ for some } x \in \Re.$$

Let a_0 and a_1 be the actions which accept, respectively, H_0 and H_1, and R a known finite measure (weight function) on $\Re$. Consider the following loss function $L(\alpha, a)$ such that

$$L(\alpha, a_0) = \int \max\{0, \tilde{F}(x) - \tilde{G}(x) - \varepsilon\} dR(x),$$

$$L(\alpha, a_1) = \int \max\{0, \varepsilon + \tilde{G}(x) - \tilde{F}(x)\} dR(x).$$

The result can be summarised in the following

Theorem 2.3.2 (Dalal and Phadia [21])

$$\Delta_n(\mathbf{X},\mathbf{Y}) = \int [\frac{\beta+n}{\beta+n+1} H^*(x',y') - \frac{\beta}{\beta+n+1} F^*(x')G^*(y')] dR(x',y').$$

2.4 Comparison of distributions

A well-known method to test the goodness of fit, based on the empirical distribution function $\hat{F}_n$, is given by the Kolmogorov-Smirnov test (see, e.g. Rohatgi [53] for details), which tests the null hypothesis

$$H_0 : F(x) = F_0(x) \ \forall x \in \Re$$

against the alternative

$$H_1 : F(x) \neq F_0(x) \text{ for some } x \in \Re.$$

The two-sided Kolmogorov-Smirnov statistics is defined as follows:

$$D_n = \sup_{x \in \Re} |\hat{F}_n(x) - F_0(x)|,$$

whereas the one-sided Kolmogorov-Smirnov statistics are defined similarly by

$$D_n^+ = \sup_{x \in \Re} \{\hat{F}_n(x) - F_0(x)\} \text{ and } D_n^- = \sup_{x \in \Re} \{F_0(x) - \hat{F}_n(x)\}.$$

The statistics D_n^+ and D_n^- are used to test H_0 against the one-sided alternatives $F(x) \geq F_0(x)$ and $F(x) \leq F_0(x)$, respectively.

When F_0 is continuous, then the exact null distributions of D_n, D_n^+ and D_n^- are known and tabulated.

Let a_0 and a_1 be the actions which accept, respectively, H_0 and H_1, and let R be a known finite measure (weight function) on $\Re^2$. Dalal and Phadia considered the loss function $L(H,a)$ such that

$$L(H,a_0) = \int \{H(x,y) - F(x)G(y)\}^- \, dR(x,y),$$

$$L(H,a_1) = \int \{H(x,y) - F(x)G(y)\}^+ \, dR(x,y),$$

where $a^+ = \max[a,0]$ and $a^- = -\min[a,0]$.

For given pairs of (vectorial) observations $(\mathbf{x},\mathbf{y})$, the probability of choosing the action a_0 is given by $\delta(x,y) = I_{\Delta_n(X,Y)\geq 0}$, where

$$\begin{aligned}\Delta_n(\mathbf{X},\mathbf{Y}) &= \mathcal{E}[L(H,a_0) - L(H,a_1)|(\mathbf{X},\mathbf{Y})] \\ &= \int \mathcal{E}[H(x',y') - F(x')G(y')|(\mathbf{X},\mathbf{Y})]dR(x',y').\end{aligned}$$

The expectation is taken w.r.t. the posterior Dirichlet process and it follows that $\mathcal{E}[H(x',y')|(\mathbf{x},\mathbf{y})]$ and $\mathcal{E}[F(x')G(y')|(\mathbf{X},\mathbf{Y})]$ are given, respectively, by

$$\frac{\beta\alpha(x',y') + \sum_{i=1}^{n} \delta_{(x_i,y_i)}\{(-\infty,x'] \times (-\infty,y']\}}{\beta+n} = H^*(x',y'),$$

and

$$\frac{p_n\alpha(x',y') + (1-p_n)\sum_{i=1}^{n} \delta_{(x_i,y_i)}\{(-\infty,x'] \times (-\infty,y']\}}{\beta+n+1} + \frac{\beta}{\beta+n+1} \cdot$$

$$\cdot \{p_n F(x') + (1-p_n)\sum_{i=1}^{n} \delta_{x_i}(-\infty,x']\} \cdot \{p_n G(y') + (1-p_n)\sum_{i=1}^{n} \delta_{y_i}(-\infty,y']\}$$

$$= \frac{H^*(x',y')}{\beta+n+1} + \frac{\beta}{\beta+n+1} F^*(x')G^*(y').$$

$$\hat{\Delta}_B = \mathcal{E}\Delta_H = \frac{\beta+n}{\beta+n+1}\left\{p_n^2\Delta_\alpha + 2p_n(1-p_n)\Delta_{(\alpha,\hat{\alpha})} + (1-p_n)^2\Delta_{\hat{\alpha}}\right\} + \frac{1}{2}\frac{1}{\beta+n+1}, \tag{2.3.1}$$

where

$$\Delta_\alpha = P_\alpha\{(X-X^{'})(Y-Y^{'})>0\} + 1/2P_\alpha\{(X-X^{'})(Y-Y^{'})=0\},$$

$$\Delta_{\hat{\alpha}} = 1/n^2\left\{\sum_{i,j=1}^{n}(I_{[(x_i-x_j)(y_i-y_j)>0]} + 1/2I_{[(x_i-x_j)(y_i-y_j)=0]})\right\},$$

$$\Delta_{(\alpha,\hat{\alpha})} = 1/n\sum_{i=1}^{n}\Delta_{\alpha(x_i,y_i)},$$

$$\Delta_{\alpha(x_i,y_i)} = P_\alpha\{(X-x_i)(Y-y_i)>0\} + 1/2P_\alpha\{(X-x_i)(Y-y_i)=0\},$$

$$p_n = \frac{\beta}{\beta+n}.$$

It is worth observing that the Bayes estimator gets close to the sample concordance coefficient as n goes to infinity, whereas it approaches the usual concordance coefficient (see Hollander and Wolfe [43]) as β goes to zero.

2.3.2 Positive dependence

Another problem dealt with by Dalal and Phadia [21] was the test of positive versus nonpositive dependence when a bivariate distribution function $H(x,y)$ is given with marginal distributions $F(x)$ and $G(y)$. We consider the null hypothesis:

$$H_0 : H(x,y) \geq F(x)G(y)\ \forall(x,y) \in \Re^2$$

against the alternative one

$$H_1 : H(x,y) < F(x)G(y)\ \forall(x,y) \in \Re^2.$$

2.3.1 Concordance

Let (X, Y) and (X', Y') be independent observations from the distribution function $H(x, y)$.

Definition 2.3.1 *The quantity*

$$\Delta = P\{(X - X')(Y - Y') > 0\} + (1/2)P\{(X - X')(Y - Y') = 0\}$$

is said to be the concordance coefficient of H.

The last term in the definition of Δ is irrelevant for continuous H; here, the term is needed because the Dirichlet process chooses continuous distribution functions with probability zero. The choice of the factor 1/2 is justified by evenly distributing the ties among concordant and discordant pairs. The concordance coefficient is used as a test of independence between X and Y, as a measure of dependence between X and Y and as a measure of concordance of observations from $H(x, y)$.

Another popular coefficient is due to Kendall [45].

Definition 2.3.2 *The quantity* $\tau = 2\Delta - 1$ *is said to be Kendall's coefficient.*

It is worth mentioning that the quantities τ and Δ become 0 and 1/2, respectively, when X and Y are independent, or H is spherically symmetric.

We consider the probability measure P_H, corresponding to the distribution function $H(x, y)$, to be a Dirichlet process on $(\Re^2, \mathcal{A})$ with parameter η and estimate the concordance coefficient, given a sample of size n and a squared loss function.

Theorem 2.3.1 (Dalal and Phadia [21]) *Let* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *be a random sample from a distribution function* $H(x, y)$. *If* H *is chosen according to a Dirichlet process* $\mathcal{D}(\eta)$, *then the Bayes estimator of the concordance coefficient* Δ_H *under the squared error loss function is given by*

Ferguson [30] proved also that, unfortunately, if $P \in \mathcal{D}(\eta)$, then P is almost surely discrete. Nonetheless, he proved that the Dirichlet process is "rich" because there is a positive probability that a sample function from P approaches any given probability measures.

In the next, we will need the following results, too. We will use α to denote both a probability measure and its distribution function, specifying which one is used when needed.

Theorem 2.2.3 (Dalal and Phadia [21]) *Let P be a Dirichlet process on $(\Re^2, \mathcal{A})$ with parameter η, and let $H(x, y)$ be the corresponding random distribution function. For any $(x, x', y, y') \in \Re^4$, it follows that*

$$\mathcal{E}H(x,y)H(x',y') = \frac{1}{\beta+1}\alpha'(x,x',y,y') + \frac{\beta}{\beta+1}\alpha(x,y)\alpha(x',y'),$$

where $\alpha'(x, x', y, y') = \alpha(x \wedge x', y \wedge y'), a \wedge b = \min\{a, b\}$.

Theorem 2.2.4 (Antoniak [3]) *Let P be a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter η, and let $B \in \mathcal{A}$. Then given $P(B) = M$, the conditional distribution of $(1/M)P$ restricted to $(B, \mathcal{A} \cap B)$ is a Dirichlet process on $(B, \mathcal{A} \cap B)$ with parameter η restricted to B. That is, if $A_1, \ldots, A_k$ is any measurable partition of B, then the conditional distribution of $(P(A_1)/M, \ldots, P(A_k)/M)$, given $P(B) = M$, is a Dirichlet distribution with parameter $(\eta(A_1), \ldots, \eta(A_k))$.*

2.3 Concordance and dependence

This section is mainly based on a paper by Dalal and Phadia [21] in which the Dirichlet process has been used to make inferences about the concordance of observations from the same distribution function.

measurable subsets $A_1, \ldots, A_m, C_1, \ldots, C_n$, *a.s.*

$$\mathcal{P}\{X_1 \in C_1, \ldots, X_n \in C_n | P(A_1), \ldots, P(A_m), P(C_1), \ldots, P(C_n)\} = \prod_{j=1}^{n} P(C_j).$$

Proposition 2.2.2 (Ferguson [30]) *Let* P *be a Dirichlet process on* $(\mathcal{X}, \mathcal{A})$ *with parameter* η*, and let* X *be a sample of size* 1 *from* P*. Then for* $B \in \mathcal{A}$*,* $\mathcal{P}(X \in B) = \alpha(B)$*.*

Let δ_x be the probability measure which assigns probability one to $x \in \mathcal{X}$.

Theorem 2.2.1 (Ferguson [30]) *Let* P *be a Dirichlet process on* $(\mathcal{X}, \mathcal{A})$ *with parameter* η*, and let* $X_1, \ldots, X_n$ *be a sample of size* n *from* P*. Then the conditional distribution of* P *given* $X_1, \ldots, X_n$ *is a Dirichlet process with parameter* $\eta + \sum_1^n \delta_{X_i}$*.*

This theorem is very important because it shows a closure property of the Dirichlet process, when going from the prior to the posterior distribution of P. In particular, theorems about Dirichlet processes apply to both priors and posteriors.

Theorem 2.2.2 (Ferguson [30]) *Let* P *be a Dirichlet process on* $(\mathcal{X}, \mathcal{A})$ *with parameter* η*, and let* Z *be a measurable real valued function defined on* $(\mathcal{X}, \mathcal{A})$*. If* $\int |Z| d\alpha < \infty$*, then* $\int |Z| dP < \infty$ *with probability one, and*

$$\mathcal{E} \int Z dP = \int Z d\mathcal{E}P = \int Z d\alpha.$$

This theorem stresses how strongly the random probability measure P depends on η. Further, when $(\mathcal{X}, \mathcal{A})$ is the real line with the Borel σ–field, and η has a finite k–th moment, then P has a finite k–th moment with probability one. The converse, although stated in Ferguson [31], was proved to be false by Doss and Sellke [29], whereas Yamato [66] provided a counterexample considering α to have a Cauchy distribution.

Finally, it is worth mentioning the following Bayes property of the Dirichlet distribution: if the prior distribution of $(Y_1, \ldots, Y_k)$ is $\mathcal{D}(\alpha_1, \ldots, \alpha_k)$ and $\mathcal{P}\{X = j | Y_1, \ldots, Y_k\} = Y_j$ a.s., for $j = 1, \ldots, k$, then the posterior distribution of $(Y_1, \ldots, Y_k)$, given $X = j$, is $\mathcal{D}(\alpha_1^{(j)}, \ldots, \alpha_k^{(j)})$, where

$$\alpha_i^{(j)} = \begin{cases} \alpha_i & i \neq j \\ \alpha_j + 1 & i = j. \end{cases}$$

2.2 The Dirichlet process

Definition 2.2.1 *Let $\mathcal{X}$ be a set and let $\mathcal{A}$ be a σ–field of subsets of $\mathcal{X}$. Let $\eta = \beta\alpha$ be a finite, nonnull, nonnegative, finitely additive measure on $(\mathcal{X}, \mathcal{A})$, with $\alpha(\mathcal{X}) = 1$ and $\beta > 0$. A random probability measure P on $(\mathcal{X}, \mathcal{A})$ is a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter η, denoted $P \in \mathcal{D}(\eta)$, if for every $k = 1, 2, \ldots$ and measurable partition $B_1, \ldots, B_k$ of $\mathcal{X}$, the joint distribution of the random probabilities $(P(B_1), \ldots, P(B_k))$ is Dirichlet with parameters $(\eta(B_1), \ldots, \eta(B_k))$, denoted $(P(B_1), \ldots, P(B_k)) \in \mathcal{D}(\eta(B_1), \ldots, \eta(B_2))$.*

As proved by Ferguson [30], this definition satisfies the Kolmogorov consistency conditions for the existence of a probability $\mathcal{P}$ on the space of all functions from $\mathcal{A}$ into $[0, 1]$ with the σ–field generated by the field of cylinder sets.

Proposition 2.2.1 (Ferguson [30]) *Let P be a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter η, and let $B \in \mathcal{A}$. If $\eta(B) = 0$, then $P(B) = 0$ with probability one. If $\eta(B) > 0$, then $P(B) > 0$ with probability one. Furthermore $\mathcal{E}P(B) = \alpha(B)$.*

Definition 2.2.2 *Let P be a random probability measure on $(\mathcal{X}, \mathcal{A})$. We say that $X_1, \ldots, X_n$ is a sample of size n from P if for any $m = 1, 2, \ldots$ and*

Definition 2.1.1 *Let $Z_1, \ldots, Z_k$ be independent random variables with $Z_i \in \mathcal{G}(\alpha_i, 1)$, with $\alpha_i \geq 0$ for all i, and $\alpha_i > 0$ for some $i, i = 1, \ldots, k$. Then the distribution of $(Y_1, \ldots, Y_k)$, with $Y_i = Z_i / \sum_{j=1}^k Z_j, i = 1, \ldots, k$, is called the Dirichlet distribution with parameter $(\alpha_1, \ldots, \alpha_k)$ and is denoted by $\mathcal{D}(\alpha_1, \ldots, \alpha_k)$.*

This distribution is always singular w.r.t. Lebesgue measure in the k–dimensional space since $Y_1 + \ldots + Y_k = 1$. In addition, Y_i is degenerate at zero if $\alpha_i = 0$. The $(k-1)$–dimensional distribution of $(Y_1, \ldots, Y_{k-1})$ is absolutely continuous if $\alpha_i > 0$ for all $i = 1, \ldots, k$, and has density

$f(y_1, \ldots, y_{k-1} | \alpha_1, \ldots, \alpha_k) =$

$$\frac{\Gamma(\alpha_1 + \ldots + \alpha_k)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} (\prod_{i=1}^{k-1} y_i^{\alpha_i - 1})(1 - \sum_{i=1}^{k-1} y_i)^{\alpha_k - 1} I_S(y_1, \ldots, y_{k-1}),$$

where $S = \{(y_1, \ldots, y_{k-1}) : y_i \geq 0, \sum_{i=1}^{k-1} y_i \leq 1\}$. If $k = 2$, then Y_1 has a Beta distribution $\mathcal{B}(\alpha_1, \alpha_2)$.

The first two moments of $(Y_1, \ldots, Y_k) \in \mathcal{D}(\alpha_1, \ldots, \alpha_k)$ are given by

$$\mathcal{E} Y_i = \alpha_i / \alpha,$$

$$\mathcal{E} Y_i^2 = \alpha_i (\alpha_i + 1) / [\alpha(\alpha + 1)],$$

$$\mathcal{E} Y_i Y_j = \alpha_i \alpha_j / [\alpha(\alpha + 1)], \text{ for } i \neq j,$$

where $\alpha = \sum_{i=1}^k \alpha_i$.

If $(Y_1, \ldots, Y_k) \in \mathcal{D}(\alpha_1, \ldots, \alpha_k)$ and $r_1, \ldots, r_l$ are integers such that $0 < r_1 < \ldots < r_l = k$, then

$$(\sum_1^{r_1} Y_i, \ldots, \sum_{r_{l-1}+1}^{r_l} Y_i) \in \mathcal{D}(\sum_1^{r_1} \alpha_i, \ldots, \sum_{r_{l-1}+1}^{r_l} \alpha_i).$$

Chapter 2

The Dirichlet Process

Bayesian nonparametric inference deals with an infinite dimensional parameter space over which a probability distribution is considered. Many approaches have been proposed since the fundamental paper by Ferguson [30] and they are reviewed by Ferguson, Phadia and Tiwari [33]. Here, we are interested in the most popular approach, based on the Dirichlet process, and we present in this chapter its properties which are going to be used in the subsequent chapters. Proofs of most of the results, along with other theorems, can be found in Ferguson [30], [31] and Antoniak [3].

2.1 The Dirichlet distribution

Let $\mathcal{G}(\alpha, \beta)$ denote the gamma distribution with shape parameter $\alpha \geq 0$ and scale parameter β. This distribution is degenerate at zero for $\alpha = 0$; otherwise, it has density with respect to (w.r.t.) Lebesgue measure on the real line:

$$f(z|\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} z^{\alpha-1} e^{-\beta z} I_{(0,\infty)}(z),$$

where $I_B(z)$ is the indicator function of the set B.

sults, on the concentration function studied by Cifarelli and Regazzini [16], by Regazzini [50], by Fortini and Ruggeri [35], [36], [37] and [38] and Ruggeri [60].

1.5 Earthquakes in Sannio Matese

We consider the interoccurrence times X (in years) of major earthquakes (i.e. with magnitudes not smaller than 5), and the number Y of minor earthquakes occurring in a given area since the previous major one. We suppose that X is exponentially distributed, while Y is geometrically distributed. Such marginals correspond to the earthquakes occurring according to a Poisson process, with constant rate, such that each earthquake has probability p of being a major one. Since such an assumption has been rejected in the analysis performed by Ladelli, Mitrione, Ruffoni and Ruggeri [46], we started using the Bayesian nonparametric approach to study relations between X and Y, finding upper and lower bounds on the concordance coefficient, showing a high level of concordance, and computing the distance, very small, between upper and lower bounds on the Bayes estimators, under squared loss function, of the random distribution.

too.

It is proved that, given a Dirichlet process P and a measurable subset B, then the distribution of the conditional distribution $P(\cdot|B)$ is a Dirichlet process, coinciding with the conditional distribution of $P(\cdot)/M$ restricted to B, given $P(B) = M$. Therefore, it is possible to find Bayes estimators, under squared loss function, of conditional probabilities and upper and lower bounds on them.

Finally, dilation is investigated comparing Bayes estimators of set probabilities and conditional set probabilities, for both prior and posterior Dirichlet processes.

1.4 Robustness and nonparametric inference

In this dissertation we present a new approach to Bayesian robustness based on nonparametrics. Whereas many studies in Bayesian robustness have dealt with a parametric family of sampling distributions, considering the prior distribution of the parameter to be uncertain, here we assume that the sampling distribution comes from a Dirichlet process with a measure $\beta\alpha$ specified with uncertainty. See Berger [4], [5], [6] and [7], and Wasserman [67] for a broad review of the results and the papers published on parametric Bayesian robustness.

In particular, we consider the distance between Dirichlet processes (based either on Hellinger distance or Kullback-Leibler divergence coefficient), the distance between distributions of set probabilities and the range spanned either by the probability of some subsets in the space of the considered probability measures or by some optimal estimators of the means or the distribution. Prior and posterior quantities are compared, too. We strongly rely, in proving re-considered, as well as the distance between their distributions. A wealth of results is contained in the book edited by Dall'Aglio, Kotz and Salinetti [25].

1.3 Parameter in a Fréchet class

Let Γ be a Fréchet class on $\Re^2$, β a positive constant and consider the class of Dirichlet processes P on $(\Re^2, \mathcal{B})$, with parameter $\eta = \beta\alpha$, where α is a probability measure in Γ. After observing a sample of size n from P, it is proved that the parameters α^* of the posterior Dirichlet processes belong to a subset Π of the posterior Fréchet class Γ^*, determined by the marginal distributions of the α^*'s. The class Ψ, containing all the distributions within upper and lower bounds on the distributions in Π, turns out to be very useful in computing bounds on quantities over Π; its properties are thus thoroughly studied. Relations between Π, Ψ and Γ^* are studied, along with distances between upper and lower bounds on the distributions contained in them, e.g. when β goes to either 0 or ∞ and when the sample size n goes to ∞.

Following a paper by Dalal and Phadia [21], one of the few on Bayesian nonparametrics on $\Re^2$, the concordance coefficient and the test of positive versus negative dependence are considered, presenting upper and lower bounds on both the Bayes estimator, under squared loss function, of the former quantity and the probability of acceptance of the null hypothesis in the latter. Bounds are computed also on the Bayes estimator of the covariance, presented in Ferguson [30].

Furthermore, a decision theoretical approach is followed by choosing, as estimates of the distribution function, the optimal ones according to some criteria (e.g. posterior regret). The consistency of the estimates is investigated,

$(\mathcal{X}, \mathcal{A})$, has been dealt with by Bayesians by considering classes of processes as priors on such a space. The Dirichlet process is the most commonly used process, having many desirable features (ease in determining the posterior probability measure, in interpreting its parameters, in computing expectations of simple loss functions, large support) with some drawbacks, not always minor, including tha fact that the process chooses a discrete probability measure with probability one. Other processes have been considered, such as neutral to the right and Polya trees, see Doksum [28] and Lavine [47] respectively, both including the Dirichlet process as a particular case, but they do not share the same tractability as the Dirichlet process. For an up to date review of the research in this field, see Ferguson, Phadia and Tiwari [33], along with the classical references to Ferguson [30], [31], Doksum [28], Antoniak [3] and Ferguson and Phadia [32]. In this dissertation, we consider the Dirichlet processes and, among other results, we give a Bayesian interpretation of the Kolmogorov-Smirnov test statistics, showing that it can be obtained from the Bayesian nonparametric statistic.

1.2 Fréchet classes

Given the univariate marginal distributions, all the probability measures on $(\Re^2, \mathcal{B})$ determine a class Γ, called the Fréchet class, which has been the object of remarkable interest in the last decades. We restrict ourselves to the space $\Re^2$, but spaces with higher dimensions can be treated similarly. The probability measures in Γ are characterised by upper and lower bounds (the so called Fréchet bounds) on their probability distribution functions. Furthermore, measures of dependence between random variables have been extensively be chosen by a Dirichlet process on $(\Re^2, \mathcal{B})$, and the knowledge, based only on the marginal probability measures, will be expressed by means of a class of Dirichlet processes with parameter proportional to a probability measure in a Fréchet class. Basic definitions and properties of Dirichlet processes and Fréchet classes are presented, respectively, in Chapter 2 and Chapter 3, along with a new, Bayesian interpretation of the Kolmogorov-Smirnov test statistics.

Properties and bounding measures of some class of parameters are studied in Chapter 4, along with some indices, like covariance and concordance, and positive dependence. The interest in such bounds and indices is justified since they are the Bayesian estimators, under squared loss function, of the random bounds and indices determined by the Dirichlet process. Chapter 4 includes also the estimation of the "optimal" distribution function according to some criteria (e.g. posterior regret), properties of a conditional probability chosen from a Dirichlet process and its application to the classes of processes here studied, and, finally, a discussion of the dilation phenomenon.

Under the uncertainty in the exact probabilistic structure, the variation of some quantities of interest is considered in Chapter 5, extending the robust Bayesian paradigm to the nonparametric case, whereas it had been previously confined to parametric inference. Finally, an application to earthquakes occurring in an Italian area called Sannio Matese is shown in Chapter 6, whereas some directions for further researches are presented in Chapter 7.

1.1 Bayesian nonparametric inference

The problem of statistical inference in an infinite dimensional parameter space, usually consisting of all the probability measures defined on a measurable space

Chapter 1

Introduction

The main goal of this dissertation consists in exploring connections between two important fields in statistical analysis which have been seldom considered together; in particular, we are interested in Bayesian nonparametric inference and the classes of probability measures with known marginals (Fréchet classes).

We consider the space of probability measures on $(\Re^2, \mathcal{B})$, where $\mathcal{B}$ is the Borel σ–field on $\Re^2$, when it is possible to make judgements only on the marginal probability measures. Such situations are common in practice. Suppose, as an example, that we want to perform inferences on two characteristics of a population which are recorded independently from each other (e.g. incomes and political opinions). We might even have indirect knowledge about the characteristics (in the previous example, we might know tax returns and political elections turnout, with all their biases). Different relations between the two characteristics are therefore allowed, by considering the Fréchet class (or a subset of it) of all the probability measures on $(\Re^2, \mathcal{B})$ sharing the same marginals.

Following the Bayesian nonparametric approach, a probability measure will

List of Tables

5.1 Hellinger distance 64

5.2 Kullback-Leibler divergence 70

6.1 Sannio Matese earthquakes data 87

6.2 Bounds on concordance 89

6.3 Distance between updated Fréchet bounds 90

D Dilation **104**

Bibliography **106**

Biography **115**

4.1.3 Relations between bounds . . . 30
4.2 Dependence between random variables . . . 36
4.2.1 Covariance . . . 36
4.2.2 Concordance . . . 37
4.2.3 Positive dependence . . . 40
4.3 Nonparametric statistical decisions . . . 41
4.3.1 Minimum expected loss criterion . . . 41
4.3.2 Minimax posterior regret criterion . . . 43
4.4 Conditional probability . . . 47
4.5 Dilation . . . 52
4.6 Distance from a given distribution . . . 55

5 Bayesian robustness **57**
5.1 Parametric Bayesian robustness . . . 57
5.2 Nonparametric Bayesian robustness . . . 59
5.3 Distance between Dirichlet processes . . . 60
5.3.1 Hellinger distance . . . 61
5.3.2 Kullback-Leibler divergence . . . 67
5.4 Distance between distributions of probabilities . . . 71
5.5 Distance between sets of probability measures . . . 74
5.6 Distance between Bayes estimators . . . 76
5.7 Comparing prior and posterior distances . . . 77

6 Earthquakes in Sannio Matese **82**
6.1 Description of the model . . . 83
6.2 Statistical analysis . . . 87
6.2.1 Fréchet class . . . 87
6.2.2 Concordance . . . 88
6.2.3 Decision . . . 89
6.2.4 Future developments . . . 91

7 Conclusions **92**

A Basic probability theorems **94**

B Concentration function and coefficients of divergence **96**

C Distribution of random functionals **101**

Contents

Abstract i

Acknowledgements iii

List of Tables vii

1 Introduction 1

1.1 Bayesian nonparametric inference 2

1.2 Fréchet classes 3

1.3 Parameter in a Fréchet class 4

1.4 Robustness and nonparametric inference 5

1.5 Earthquakes in Sannio Matese 6

2 The Dirichlet Process 7

2.1 The Dirichlet distribution 7

2.2 The Dirichlet process 9

2.3 Concordance and dependence 11

2.3.1 Concordance 12

2.3.2 Positive dependence 13

2.4 Comparison of distributions 15

3 Fréchet classes 19

3.1 Definition and main properties 19

4 Dirichlet process and Fréchet class 22

4.1 Updating the class of parameters 22

4.1.1 Classes of probability measures 23

4.1.2 Expectations 30

Acknowledgements

This dissertation is the final step of my career as a student: it has been one of the longest career I have never heard of, with some long interruptions in it. Therefore, trying to remember all those who made possible this achievement of mine is quite difficult, so I explicitly acknowledge only the people who contributed the most, thanking all together the people at Università di Milano, at Carnegie Mellon University and Duke University who were close to me during my studies, and the friends whom I had the pleasure to work with in the last years.

I wish to thank the members of my committee, especially Michael Lavine who has been a very stimulating and demanding advisor. My thanks go to Mike West, always ready to help with scientific and practical problems.

Another important place is reserved to my colleagues at the Istituto per le Applicazioni della Matematica e dell'Informatica del Consiglio Nazionale delle Ricerche, in particular to the Director, Eugenio Regazzini, and my first advisor, Bruno Betrò.

But the last, most important acknowledgements are due to my family: we know what I owe them.

cess, then it follows that the conditional probabilities, conditioned on a given subset, are still realisations of a suitable Dirichlet process. It is therefore possible to find Bayes estimators, under squared loss function, of conditional probabilities and upper and lower bounds on them.

It is shown that dilation occurs when comparing Bayes estimators of set probabilities and conditional set probabilities, for both prior and posterior Dirichlet processes.

A new, nonparametric approach to Bayesian robustness is introduced, considering some distances between Dirichlet processes, the distance between distributions of set probabilities and the range spanned either by the probability of some subsets in the space of the probability measures or by some optimal estimators of the means or the distribution. Prior and posterior quantities are compared, too.

Finally, an application of some results to earthquakes occurring in an Italian area called Sannio Matese is shown.

Abstract

Connections between Bayesian nonparametric inference and the classes of probability measures with known marginals (Fréchet classes) are investigated. In particular, a class of Dirichlet processes with parameters proportional to probability measures in a Fréchet class will be considered.

Relations between two statistical characters are then studied, by presenting a new, Bayesian interpretation of the Kolmogorov-Smirnov test statistics, by looking for upper and lower bounds on Bayes estimators, under squared loss function, of some indices, like covariance and concordance, and by considering positive dependence and "optimal" estimators of distribution functions according to some criteria (e.g. posterior regret). It should be pointed out that, in general, bounds are found when considering Dirichlet processes whose parameter is either one of the prior Fréchet bounds or one obtained by updating them after observing a sample.

Properties of some classes of probability measures are investigated too, along with their relations. In general, computing upper and lower bounds in a suitable larger class turns out to be easier than in the true class, whereas it is shown that the bounds on the two classes coincide.

It is proved that, given a probability measure chosen from a Dirichlet pro-

<u>ABSTRACT</u>

(Statistics)

BAYESIAN NONPARAMETRICS, ROBUSTNESS AND FRÉCHET CLASSES

by

Fabrizio Ruggeri

Institute of Statistics and Decision Sciences
Duke University

Date: ____________________
Approved:

____
Dr. Michael Lavine, Supervisor

____
Dr. Peter Mueller

____
Dr. Ken Reckhow

____
Dr. Brani Vidakovic

____
Dr. Robert Wolpert

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the Institute of
Statistics and Decision Sciences in the Graduate School of
Duke University

1994

BAYESIAN NONPARAMETRICS, ROBUSTNESS AND FRÉCHET CLASSES

by

Fabrizio Ruggeri

Institute of Statistics and Decision Sciences
Duke University

Date: ______________________________

Approved:

____
Dr. Michael Lavine, Supervisor

____
Dr. Peter Mueller

____
Dr. Ken Reckhow

____
Dr. Brani Vidakovic

____
Dr. Robert Wolpert

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Institute of Statistics and Decision Sciences in the Graduate School of Duke University

1994

Copyright © 1995 by Fabrizio Ruggeri
All rights reserved