Modeling Statistical Dependence

Chong Gu

June 28, 2006

The modeling of statistical dependence is a vast subject area which I do not even dream to cover comprehensively. In this discussion, I try to give a brief overview of the corner of the universe I personally have traversed a bit. The specific subjects I will touch upon include the graphical models for conditional independence and some issues related to correlated errors in regression models.

1 Graphical Models

Consider random variables X, Y, Z, ... The (conditional) dependence structures can be represented by graphs, and models exploring/exploiting such structures are called graphical models.

1.1 Functional ANOVA and (Conditional) Independence

Take the density $f(x,y)$ of (X,Y), and write $f(x,y) = e^{\eta(x,y)}/\int_{\mathcal{X}\times\mathcal{Y}} e^{\eta(x,y)}$. With a functional ANOVA decomposition, $\eta(x,y) = \eta_\emptyset + \eta_x(x) + \eta_y(y) + \eta_{x,y}(x,y)$, where η_x, η_y, and $\eta_{x,y}$ satisfy side conditions such as $\int_{\mathcal{X}} \eta_x = \int_{\mathcal{X}} \eta_{x,y} = \int_{\mathcal{Y}} \eta_y = \int_{\mathcal{Y}} \eta_{x,y} = 0$, one sets $\eta_\emptyset = 0$ for a one-to-one logistic density transform. The independence of X and Y, denoted by $X \perp Y$, is characterized by $\eta_{x,y} = 0$.

For (X,Y,Z), one has $f(x,y,z) = e^{\eta(x,y,z)}/\int e^{\eta(x,y,z)}$, where $\eta = \eta_x + \eta_y + \eta_z + \eta_{x,y} + \eta_{x,z} + \eta_{y,z} + \eta_{x,y,z}$. The conditional density of $(Y,Z)|X$ is given by

$$f(y,z|x) = \frac{e^{\eta_y+\eta_z+\eta_{x,y}+\eta_{x,z}+\eta_{y,z}+\eta_{x,y,z}}}{\int_{\mathcal{Y}\times\mathcal{Z}} e^{\eta_y+\eta_z+\eta_{x,y}+\eta_{x,z}+\eta_{y,z}+\eta_{x,y,z}}}$$

The conditional independence of Y and Z given X, denoted by $(Y \perp Z)|X$, is characterized by $\eta_{y,z} + \eta_{x,y,z} = 0$.

Note that X, Y, Z here are generic. In particular, they can each be random vectors themselves, and the functional ANOVA can be defined recursively. For example, with $X = (U,V)$, one has $\eta_x = \eta_u + \eta_v + \eta_{u,v}$ and $\eta_{x,y,z} = \eta_{u,y,z} + \eta_{v,y,z} + \eta_{u,v,y,z}$. If the density of (U,V,Y,Z) has the expression

$$f(u,v,y,z) = \frac{e^{\eta_u+\eta_v+\eta_y+\eta_z+\eta_{u,y}+\eta_{u,z}+\eta_{v,y}+\eta_{v,z}}}{\int_{\mathcal{U}\times\mathcal{V}\times\mathcal{Y}\times\mathcal{Z}} e^{\eta_u+\eta_v+\eta_y+\eta_z+\eta_{u,y}+\eta_{u,z}+\eta_{v,y}+\eta_{v,z}}},$$

then $(U \perp V)|(Y,Z)$ and $(Y \perp Z)|(U,V)$.

Also note that the elimination of terms beyond the second order does *not* induce independence structures, although it makes estimation easier.

1.2 Gaussian Models

For Gaussian distribution, the log density η is a quadratic form, so only the first and second order terms are present. Conditional independence structures can be characterized by 0's in the coefficient matrix, the inverse of the variance-covariance matrix.

On finite domains (truncated normal?), the coefficient matrix does not have to be non-negative definite, but its inverse is no longer variance-covariance.

1.3 Log Linear Models

Many problems in categorical data analysis can be approached via surrogate Poisson regression models, better known as log linear models, which form interesting special cases of graphical models. In fact, modern graphical models are largely motivated by classical log linear models.

Consider a two-way contingency table with entries $n_{i,j}$. Assuming n_{ij} as from Poisson(λ_{ij}), one may write

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij},$$

where α_i, β_j, $(\alpha\beta)_{ij}$ satisfy the usual ANOVA side conditions. The independence of the two margins is characterized by $(\alpha\beta)_{ij} = 0$.

Similarly, for a three-way table with entries $n_{ijk} \sim \text{Poisson}(\lambda_{ijk})$, one has

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk},$$

and the conditional independence of the j and k margins given the i margin is characterized by $(\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} = 0$.

Note that the (i, j, k) here are the (X, Y, Z) in §1.1, and $\lambda_{i,j,k}$ is effectively $f(x, y, z)$ on the discrete product domain but *without* being normalized.

Poisson regression is equivalent to density estimation up to the normalizing constant, on any domain.

1.4 Conditional Gaussian Distributions

Much of modern research on graphical models has been focusing on the development/characterization of parametric distributions for mixtures of continuous and discrete random variables. These distributions are called Conditional Gaussian (CG) as the conditional distributions of the continuous random variables given the discrete ones are Gaussian.

1.5 Nonparametric Graphical Models via Penalized Likelihood

It is rather challenging to develop parametric graphical models for mixtures of continuous and discrete variables, or even for purely continuous variables besides the Gaussian models. On the other hand, it is straightforward to characterize graphical models via functional ANOVA decomposition on any product domain, assuming one can estimate the log density η nonparametrically.

An approach to nonparametric function estimation is the penalized likelihood method. Consider an univariate regression problem with $Y_i = \eta(x_i) + \epsilon_i$, where $x_i \in [0, 1]$ and $\epsilon_i \sim N(0, \sigma^2)$. The method estimates η through the minimization of

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \eta(x_i))^2 + \lambda \int_0^1 \ddot{\eta}^2, \tag{1}$$

which yields the cubic smoothing (natural) splines.

For the estimation of probability density $f(x) = e^{\eta(x)}/\int_{\mathcal{X}} e^{\eta(x)}$ on domain $\mathcal{X}$ using independent samples X_i, one may minimize

$$-\frac{1}{n}\sum_{i=1}^{n}\Big\{\eta(X_i) - \log\int_{\mathcal{X}} e^{\eta(x)}\Big\} + \frac{\lambda}{2}J(\eta).$$

Note that the domain $\mathcal{X}$ is generic, which can be a product domain of arbitrary marginals. $J(\eta)$ is a quadratic roughness functional and the minimization takes place in a so-called reproducing kernel Hilbert space, say $\mathcal{H}$. On product domains, $\mathcal{H}$ and $J(\eta)$ can be constructed with functional ANOVA decomposition built in.

For (X, Y) with joint density $f(x, y) \propto e^{\eta_x+\eta_y+\eta_{x,y}}$ on $\mathcal{X} \times \mathcal{Y}$, the conditional density of $Y|X$ is seen to be $f(y|x) = e^{\eta_y+\eta_{x,y}}/\int_{\mathcal{Y}} e^{\eta_y+\eta_{x,y}}$. To estimate $\eta(x, y) = \eta_y(y) + \eta_{x,y}(x, y)$ from (X_i, Y_i), one may minimize

$$-\frac{1}{n}\sum_{i=1}^{n}\Big\{\eta(X_i, Y_i) - \log\int_{\mathcal{Y}} e^{\eta(X_i,y)}\Big\} + \frac{\lambda}{2}J(\eta).$$

Again note that $\mathcal{X}$ and $\mathcal{Y}$ are generic here, so the simple formulation yields a rich collection of versatile statistical models.

1.6 References

A quick, entertaining reading on graphical models with primarily Gaussian examples is found here.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.

Log linear models are found in many books including those by McCullagh and Nelder and by Agresti. Many years ago I enjoyed reading the following book, but all details are forgotten now.

Plakett, R. L. (1974). *The Analysis of Categorical Data*. London: Griffin.

Among names appearing frequently in the modern graphical model literature, especially concerning the CG distributions, are Edwards, Lauritzen, and Wermuth. There are a few books published including this one, but I did not read any.

Lauritzen, S. L. (1996). *Graphical Models*. New York: Oxford University Press.

Also, I have the following book that I received for reviewing book proposals for the publisher, which I have yet to read.

Cox D. R. and Wermuth, N. (1996). *Multivariate Dependencies: Models, analysis and interpretation*. New York: Chapman and Hall/CRC.

The materials of §1.1 and §1.5 are taken from my book, Section 1.3 and Chapter 6.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. New York: Springer-Verlag.

Graphical models are more for the modeling of conditional independence rather than the characterization of dependence. For the modeling/characterization of statistical dependence, different concepts/tools are needed. Among important issues is the ordering of dependence, and among important modeling tools are copulas. These and many other related topics can be found in the following book.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. New York: Chapman and Hall/CRC.

2 Regression with Correlated Errors

Consider a regression problem with $Y_i = \eta(x_i) + \tilde{\epsilon}_i$, where $\tilde{\boldsymbol{\epsilon}} \sim N(\mathbf{0}, \sigma^2 W^{-1})$. The estimation of $f(x)$ is often of primary interest, whereas W may sometimes be a nuisance. Typically, one needs to specify parametric models for the correlation W.

2.1 Random/Mixed Effect Models, Longitudinal Data

For the analysis of longitudinal data, clustered observations, or the like, a standard modeling technique is to decompose ϵ into the sum of random effects and independent measurement errors,

$$\tilde{\epsilon}_i = \mathbf{z}_i^T \mathbf{b} + \epsilon_i,$$

where $\mathbf{b} \sim N(\mathbf{0}, B)$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \tau^2 I)$. It is easily seen that $\sigma^2 W^{-1} = ZBZ^T + \tau^2 I$, where $Z^T = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ and B and τ^2 are to be specified/estimated.

With $Y_i = \eta(x_i) + \mathbf{z}_i^T \mathbf{b} + \epsilon_i$, $\eta(x_i)$ is the fixed effect, $\mathbf{z}_i^T \mathbf{b}$ is the random effect, and the model is known as a mixed-effect model. Focusing on the structure of the error variance-covariance $\sigma^2 W^{-1}$, the model is also called a variance component model with variance components ZBZ^T and $\tau^2 I$; ZBZ^T could often be further decomposed.

Consider longitudinal observations Y_i taken from subject $s_i \in \{1, \ldots, p\}$ with covariate x_i, a simple model for intra-subject correlation is given by

$$Y_i = \eta(x_i) + b_{s_i} + \delta_i$$

with $b_s \sim N(0, \sigma_b^2)$; here $\mathbf{b}^T = (b_1 \ldots, b_p)$, $B = \sigma_b^2 I$, and $\mathbf{z}_i = \mathbf{e}_{s_i}$ with $\mathbf{e}_s$ the sth unit vector.

Simple adaptations of this technique can be used to model correlations in other settings. For example, with $Y_i \sim \text{Binomial}(m_i, p_i)$, one may model the logit $\eta = \log\{p/(1-p)\}$ via

$$\eta_i = \eta(x_i) + \mathbf{z}_i^T \mathbf{b}.$$

For hazard estimation using lifetime data, one may add $\mathbf{z}_i^T \mathbf{b}$ to the log hazard to obtain the frailty models. The computation in non-Gaussian settings is non-trivial.

2.2 Sequential/Spatial Correlation

Sequential correlation is the subject of time series analysis. Spatial correlation is the "natural" extension of sequential correlation, and a common model is the AR(1)-type, $E[\tilde{\epsilon}(t)\tilde{\epsilon}(s)] = \sigma^2 e^{-\alpha|t-s|}$, where α is positive and $|t-s|$ is the Euclidean distance between locations t and s; sometimes a further independent measurement error is added on top of this.

For a clean model identifiability the covariate x_i should be "orthogonal" to the time line or the spatial location, although a parametric $\eta(x_i)$ could be safe in any case.

The random-effect/variance-component models of §2.1 usually result in low-rank modifications of the covariance matrix. The sequential/Spatial correlation models typically yield full-rank "modifications."

Tricky to adapt for non-Gaussian data.

2.3 Penalty Smoothing and Empirical Bayes

Recall the cubic smoothing spline as the minimizer of (1). The solution $\eta_\lambda(x)$ coincides with an empirical Bayes estimate under a certain prior for $\eta(x)$, in the sense that $E[\eta(x)|\mathbf{Y}] = \eta_\lambda(x)$, $\forall x$.

The prior that leads to this mathematical equivalence has two independent components, $\eta = \eta_0 + \eta_1$, where η_0 is diffuse in the null space span$\{1, x\}$ of $J(\eta) = \int \ddot{\eta}^2$, and η_1 has a mean 0 Gaussian process prior with a covariance function $E[\eta_1(x_1)\eta_1(x_2)] = bR(x_1, x_2)$, where $b = \sigma^2/n\lambda$ and R is the "inverse" of $\int \ddot{\eta}^2$. In more familiar terms, one may write

$$Y_i = \beta_0 + \beta_1 x_i + \eta_1(x_i) + \epsilon_i, \tag{2}$$

where $\beta_0 + \beta_1 x_i$ is the "parametric" fixed effect and $\eta_1(x_i)$ is the "nonparametric" random effect.

In general, the minimizer of the penalized least squares functional

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \eta(x_i))^2 + \lambda J(\eta)$$

is always an empirical Bayes solution on any domain, with fixed effect in the null space of $J(\eta)$ and random effect with the covariance proportional to the "inverse" of $J(\eta)$.

2.4 Spatial Correlation Models

Most of the spatial correlation models used in geostatistics/kriging are in forms similar to (2), thus can be perceived as penalty smoothing, except that $R(x_1, x_2)$ would be more intuitive whereas the corresponding $J(\eta)$ may not have an explicit expression.

There seems to be a fundamental identifiability problem between a "smooth" fixed effect and a "nonparametric" random effect, so try not to incorporate both in the same model. The "nonparametric" random effect is used to model the signal here, *not* to model the noise as in §2.2.

Parameters in the covariance $R(x_1, x_2)$ may not always be identifiable/estimable. [Equivalence and perpendicularity of probability measures.]

2.5 Smoothing with Correlated Errors

To estimate $\eta(x)$ nonparametrically, one approach is to use the penalized likelihood method via the minimization of

$$(\mathbf{Y} - \boldsymbol{\eta})^T W (\mathbf{Y} - \boldsymbol{\eta}) + n\lambda J(\eta),$$

where $\boldsymbol{\eta}^T = (\eta(x_1), \ldots, \eta(x_n))$.

For W completely known there is little difference between this and penalized least squares. For W known only up to a few parameters one needs to select/estimate the correlation parameters along with the smoothing parameter λ.

2.6 Asymptotic Convergence?

For smoothing with independent errors, one may assume a limiting density $f(x)$ for the covariate and calculate the asymptotic convergence rates in terms of $\int(\hat{\eta}(x) - \eta(x))^2 f(x)dx$. With correlated data, however, the plain, unweighted mean square error may not be a reasonable cause to pursue, but how can one incorporate the dependence structure into an asymptotic analysis?

2.7 References

Mixed-effect models are widely used in applications. Influential writings/software for parametric models include those by Laired and by Bates. For interesting perceptions/insights I like this paper.

Robinson, G. K. (1991). "That BLUP is a good thing: The estimation of the random effects," *Statist. Sci. 6*, 15–51 (with discussions).

For technical results I look up this one.

Harville, D. A. (1977). "Maximum likelihood approaches to variance component estimation and to related problems," *J. Amer. Statist. Assoc. 72*, 320–340 (with discussions).

Sequential correlation models are the subjects of time series analysis on which many books are available. Though a bit dated, I still find the following volume a handy reference.

Priestley, M. B. (1981). *Spectral Analysis and Times Series*. London: Academic Press.

The equivalence of penalty smoothing with quadratic penalties and empirical Bayes with Gaussian process priors was observed no later than the early 1970s by Kimeldorf and Wahba. I would look up Wahba's book for the authentic rendering.

Wahba, G. (1990). *Spline Models for Observational Data*, Volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: SIAM.

For the equivalence and perpendicularity of probability measures Wahba quoted Shepp. Stein showed how "misspecified" covariance may not necessarily hurt prediction.

Smoothing with correlated errors was the subject of two dissertations I supervised, by Ping Ma in 2003 and by Chun Han in 2005.