SIMPLEX REGRESSION

MARK FISHER

ABSTRACT. This note characterizes a class of regression models where the set of coefficients is restricted to the simplex (i.e., the coefficients are nonnegative and sum to one). This structure arrises in the context of fitting a functional form nonparametrically where the functional form is subject to shape constraints of a particular sort. Two examples are given. The approach to inference is Bayesian, using a Dirichlet-based sparsity prior. A variety of approaches to sampling from the posterior distribution are presented.

Date: May 27, 2016 @ 07:12. *Filename*: "simplex regression".

JEL Classification. C11, C14, G13.

Key words and phrases. Bayesian nonparametric regression, simplex, implied probability distributions, futures and options.

The views expressed herein are the author's and do not necessarily reflect those of the Federal Reserve Bank of Atlanta or the Federal Reserve System.

1. Introduction

This note characterizes a class of regression models where the set of coefficients is restricted to the simplex (i.e., the coefficients are nonnegative and sum to one). Two examples are presented that suggest the need for an under-identified setup. In the examples, the coefficients themselves are nuisance parameters: They are part of the framework that guarantees certain restrictions are maintained and they are not of interest on their own.

The approach to inference is Bayesian, using a Dirichlet-based sparsity prior. A variety of approaches to sampling from the posterior distribution are presented.

Section 2 presents the model and the likelihood. Section 3 presents two examples. Section 4 presents the prior. Sections 5 and 6 present a Gibbs sampler and an importance sampler, respectively.

2. Model

The class of regression models under consideration here are characterized by the following:

$$y_i = \lambda \sum_{j=1}^{K} X_{ij}\, \beta_j + \varepsilon_i, \tag{2.1}$$

where $\lambda > 0$, $\varepsilon_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma_\varepsilon^2)$, and $\beta = (\beta_1, \ldots, \beta_K) \in \Delta^{K-1}$, where Δ^{K-1} denotes the simplex of dimension $(K-1)$. In other words, β satisfies $\sum_{j=1}^{K} \beta_j = 1$ and $\beta_j \geq 0$ for $j = 1, \ldots, K$.

Let $y = (y_1, \ldots, y_N)$. Given (2.1), one can stack the observations in vector form:

$$y = \lambda X \beta + \varepsilon, \tag{2.2}$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_N) \sim \mathsf{N}(0_N, \sigma_\varepsilon^2\, I_N)$. Note that X is an $N \times K$ matrix of observed covariates. One can express (2.2) as a likelihood for the unobserved parameters:

$$p(y|\lambda, \beta, \sigma_\varepsilon^2) = \mathsf{N}(y|\lambda X\beta, \sigma_\varepsilon^2\, I_N) = (2\,\pi\,\sigma_\varepsilon^2)^{-N/2} \exp\left(\frac{-S(\lambda,\beta)}{2\,\sigma_\varepsilon^2}\right), \tag{2.3}$$

where[1]

$$S(\lambda, \beta) := (y - \lambda X\beta)^\top (y - \lambda X\beta). \tag{2.4}$$

The examples below suggest the desirability of $K > N$, in which case β will not be fully identified. The goal will be to average across reasonable values for β.

Discussion. It is simple to impose the adding-up restriction by letting $\beta_k = 1 - \sum_{j \neq k} \beta_j$ for some $k \in \{1, \ldots, K\}$ and rewriting (2.1) as

$$y_i - \lambda X_{ik} = \lambda \sum_{j \neq k} \beta_j\, (X_{ij} - X_{ik}) + \varepsilon_i. \tag{2.5}$$

To express this more compactly, let X_{*k} denote the k-th column of X and subtract λX_{*k} from both sides of (2.2) to produce

$$y_\lambda^k = \lambda X^k \beta + \varepsilon, \tag{2.6}$$

[1] $A^\top$ denotes the transpose of A.

where

$$y^k_\lambda := y - \lambda X_{*k} \tag{2.7}$$

and where the j-th column of the matrix X^k is given by

$$X^k_{*j} := X_{*j} - X_{*k}. \tag{2.8}$$

Although β_k appears in (2.6) explicitly, the k-th column of X^k is the zero vector and consequently β_k vanishes. The non-negativity constraints can be dealt with via a posterior sampler that apprehends $\Pr[\beta \notin \Delta^{K-1}] = 0$.

Taking a different approach, all of the restrictions on β can be imposed via a reparametrization. Let $v = (v_1, \ldots, v_K)$ and let

$$\beta_j = \frac{e^{v_j}}{\sum_{j=1}^K e^{v_j}}. \tag{2.9}$$

We may express this dependence as $\beta(v)$. Then $v \in \mathbb{R}^K \implies \beta(v) \in \Delta^{K-1}$. Thus v is unrestricted in $\mathbb{R}^K$. However, note that if $v' = (v_1 + c, \ldots, v_K + c)$, then $\beta(v') = \beta(v)$. Consequently, even if β is identified, v will not be. A prior for v will reduce or remove the implicit indeterminacy.

Prior and posterior. Given a prior distribution $p(\lambda, \beta, \sigma^2_\varepsilon)$ for the unknown parameters, the posterior distribution can be expressed as

$$p(\lambda, \beta, \sigma^2_\varepsilon | y) \propto p(y|\lambda, \beta, \sigma^2_\varepsilon)\, p(\lambda, \beta, \sigma^2_\varepsilon). \tag{2.10}$$

I present priors for the unknown parameters in Section 4.[2] The marginal posterior distribution for β is given by

$$p(\beta|y) = \iint p(\lambda, \beta, \sigma^2_\varepsilon | y)\, d\sigma^2_\varepsilon\, d\lambda. \tag{2.11}$$

3. Two examples

I present two examples that have the form of simplex regression. Both examples are related to probability distributions.

First example. Consider the problem of inferring the risk-neutral probability density from put option data.[3] The observed value of a European derivate security with payout function $g_i(x)$ is

$$y_i = B \int_{-\infty}^{\infty} g_i(x)\, q(x)\, dx + \varepsilon_i, \tag{3.1}$$

where $q(x)$ is the *risk-neutral probability* distribution for the *underlier*, B is the *discount factor* (i.e., the value of a risk-free discount bond that matures on the expiration date), and ε_i is "measurement error" (of one sort or another; more on this below).

For a put option, $g_i(x) = \max[s_i - x, 0]$, where s_i is the *strike price.* Based on this payout function, define

$$v(s) := B \int_{-\infty}^{\infty} \max[s - x, 0]\, q(x)\, dx = B \int_{-\infty}^{s} (s - x)\, q(x)\, dx. \tag{3.2}$$

[2] The priors will involve a hyperparameter which is omitted here for expositional simplicity.

[3] Call option data can easily be incorporated as well.

The observed value of a put option can be expressed as

$$y_i = v(s_i) + \varepsilon_i. \tag{3.3}$$

Note that v is subject to the following shape and boundary restrictions:

$$v(s) \geq 0, \qquad v'(s) \geq 0, \qquad \text{and} \qquad v''(s) \geq 0 \tag{3.4}$$

and

$$\lim_{s\to-\infty} v(s) = 0, \qquad \lim_{s\to-\infty} v'(s) = 0, \qquad \text{and,} \qquad \lim_{s\to\infty} v'(s) = B. \tag{3.5}$$

In particular, note that $v''(s) \equiv B\,q(s)$. From this perspective it is natural to view the problem of inferring the risk-neutral density as an exercise in nonparametric regression subject to a set of shape constraints.[4]

A complementary approach is to directly represent $q(x)$ nonparametrically in such a way as to automatically satisfy all of the shape restrictions. In order to represent the unknown density flexibly, one can adopt a mixture of *basis densities*. Basis densities are basis functions that satisfy conditions that guarantee they are valid densities. Consider a collection of basis densities $\{f_j(x)\}_{j=1}^K$, where $\int_{-\infty}^{\infty} f_j(x)\,dx = 1$ and $f_j(x) \geq 0$ for $x \in (-\infty,\infty)$. See the Appendix for the description of a useful class of basis densities.

Define

$$f(x|\beta) := \sum_{j=1}^{K} \beta_j\, f_j(x). \tag{3.6}$$

Replacing $q(x)$ with $f(x|\beta)$, we have

$$\int_{-\infty}^{\infty} g_i(x)\, f(x|\beta)\,dx = \sum_{j=1}^{k} \beta_j\, X_{ij}, \tag{3.7}$$

where

$$X_{ij} = \int_{-\infty}^{\infty} g_i(x)\, f_j(x)\,dx. \tag{3.8}$$

Substituting (3.7) into (3.1) and letting $\lambda = B$ produces (2.1). From this perspective, ε_i may be understood as both measurement error and model error, the latter resulting from the possibility that $q(x)$ lies outside the space spanned by the basis densities.

In passing, note that point masses (as represented by Dirac delta functions) can be used as basis densities: $f_j(x) = \delta(x - m_j)$, where m_j is the location of the j-th basis density and $\delta(\,\cdot\,)$ is the Dirac delta function.[5] Then [referring to (3.8)]

$$X_{ij} = \textstyle\int_{-\infty}^{\infty} g_i(x)\,\delta(x - m_j)\,dx = g_i(m_j) = \max[s_i - m_j, 0].$$

Carlson et al. (2005) implicitly adopt these basis densities. They do not, however, undertake a Bayesian approach to estimation.

By the nature of the problem posed, a suitably flexible set of basis densities may require $K > N$ which will lead to under-identification. The prior distribution and the posterior

[4] There is a substantial literature that deals with the problem from this perspective. For example, see Aït-Sahalia and Duarte (2003).

[5] The two relevant properties of the delta function are $\int_{-\infty}^{\infty} \delta(x - x_0)\,dx = 1$ and $\int_{-\infty}^{\infty} f(x)\,\delta(x - x_0)\,dx = f(x_0)$.

samplers described below are chosen with this this aspect of the problem in mind. In particular, the sparsity prior presented below allows for the stable estimation of a large number of mixture coefficients.

As noted in the introduction, the β vector is not itself of interest. Rather, it is the function $f(x|\beta)$ — accounting for the uncertainty regarding β — that is of interest. To this end, one integrates out β using its posterior distribution $p(\beta|y)$ to obtain

$$\widehat{f}(x) := \int f(x|\beta)\, p(\beta|y)\, d\beta = \sum_{j=1}^{K} \widehat{\beta}_j\, f_j(x) = f(x|\widehat{\beta}), \tag{3.9}$$

where

$$\widehat{\beta} = E[\beta|y]. \tag{3.10}$$

This superficially resembles a "plug-in" estimator. However, note that $\widehat{\beta}$ is determined by integration and not by optimization. In any event, the latter route may not be available owing to under-identification.

Second example. Consider the problem of fitting a cumulative distribution function (CDF) to observed data of the following form:[6] $\{(s_i, y_i)\}_{i=1}^n$, where the s_i are distinct and

$$y_i = \Pr[x \le s_i]. \tag{3.11}$$

Let $G(s) = \int_{-\infty}^{s} g(x)\, dx$, where $g(x)$ is an unknown density. Note that $G(\infty) = 1$. The possibility of a point mass at infinity can be accomodated by expressing the relation between s_i and y_i as follows:

$$y_i = (1 - w)\, G(s_i) + \varepsilon_i, \tag{3.12}$$

where $w \in [0, 1)$. The point mass can be eliminated by setting $w = 0$.

Now consider a collection of basis distributions $\{F_j(x)\}_{j=1}^K$, where $F_j'(x)$ is a basis density.[7] Define

$$F(x|\beta) := \sum_{j=1}^{K} \beta_j\, F_j(x). \tag{3.13}$$

Replacing $G(x)$ with $F(x|\beta)$, we have

$$F(s_i|\beta) = \sum_{j=1}^{K} \beta_j\, X_{ij}, \tag{3.14}$$

where $X_{ij} = F_j(s_i)$. Letting $\lambda = (1 - w)$, we can express (3.12) as (2.1).

4. Prior distributions

In this section I describe the prior distributions for the unknown parameters $(\sigma_\varepsilon^2, \lambda, \beta, \alpha)$.

[6] See Fisher (2015) for a fleshed-out example that involves the probability distribution for the half-life of deviations from purchasing power parity.

[7] See the Appendix for an example of basis distributions.

Prior for σ_ε^2. For the purpose at hand, σ_ε^2 is a nuisance parameter. As such, I adopt a Jeffreys prior:

$$p(\sigma_\varepsilon^2) \propto 1/\sigma_\varepsilon^2. \tag{4.1}$$

Prior for λ. Three priors for λ will be entertained. First, one may assume λ is known (i.e., a dogmatic prior); for example, $\lambda = 1$. Second, one may assume λ is normally distributed (truncated at zero). And third, one may assume the improper prior $p(\lambda) \propto 1_{(0,\infty)}(\lambda)$, which we may interpret as the limit of a truncated normal as the variance goes to infinity.

Prior for β. Now we focus on the prior for β.

A benefit of the Bayesian approach is the ability to incorporate important considerations via the prior for β. In particular, the prior for β should embody two features. First the prior should ensure the constraint $\beta \in \Delta^{K-1}$, where Δ^{K-1} is the $(K-1)$-dimensional simplex. Second the prior should be capable of expressing the idea that while any of the basis densities is possible, only a few should have nontrivial probability associated with it. The Dirichlet distribution embodies both of these features. Let

$$p(\beta|\alpha) = \mathsf{Dirichlet}(\beta|\alpha\,\xi) = \frac{\Gamma(\alpha)}{\prod_{j=1}^K \Gamma(\alpha\,\xi_j)} \prod_{j=1}^K \beta_j^{\alpha\,\xi_j - 1}, \tag{4.2}$$

where $\alpha > 0$, $\xi \in \Delta^{K-1}$, and $\alpha\,\xi = (\alpha\,\xi_1, \ldots, \alpha\,\xi_K)$. Note $E[\beta|\alpha,\xi] = \xi$. I will refer to α as the *concentration parameter*. If $\xi_j = 1/K$ and $\alpha = K$, then the prior is flat: $p(\beta|\alpha) = (K-1)!$.

Prior for α. As K gets large relative to n, the flat prior for β tends to dominate the likelihood (2.3) and the posterior can become quite flat itself. Setting $\alpha < K$ encourages parsimony, pushing the mass of the prior toward the vertices of the simplex, thereby implicitly suggesting that only a few of the components are nonnegligible. This prior could be described as "partially informed ignorance."

On the other hand, if ξ is well-informed (because, for example, it is based on closely-related data) then $\alpha > K$ may be suitable. I provide a prior for α that encourages sparsity, but also allows the data to overrule the prior and emphasize other values for α with more or less concentrated posterior distributions.

Let $z = \log(\alpha)$ and let

$$p(z) = \mathsf{N}(z|\zeta, \tau^2), \tag{4.3}$$

where ζ and τ are the mean and standard deviation parameters (respectively). The prior for z implies the following prior for α:

$$p(\alpha) = \frac{\mathsf{N}(\log(\alpha)|\zeta, \tau^2)}{\alpha} = \mathsf{Log\text{-}Normal}(\alpha|\zeta, \tau^2). \tag{4.4}$$

The prior median for α is e^ζ.

5. First posterior sampler

This section describes a Gibbs sampler. Given the likelihood and the prior, the posterior distribution for the parameters can be expressed as

$$p(\lambda, \beta, \sigma_\varepsilon^2, \alpha|y) \propto \frac{p(y|\lambda, \beta, \sigma_\varepsilon^2)\, p(\beta|\alpha)\, p(\lambda)\, p(\alpha)}{\sigma_\varepsilon^2}. \tag{5.1}$$

The Gibbs sampler cycles through the following full conditional posterior distributions:

$$p(\sigma_\varepsilon^2|y, \lambda, \beta, \alpha) \tag{5.2a}$$

$$p(\lambda|y, \sigma_\varepsilon^2, \beta, \alpha) \tag{5.2b}$$

$$p(\beta|y, \lambda, \sigma_\varepsilon^2, \alpha) \tag{5.2c}$$

$$p(\alpha|y, \lambda, \sigma_\varepsilon^2, \beta). \tag{5.2d}$$

Drawing σ_ε^2. Drawing from the conditional posterior for σ_ε^2 is straightforward. Note

$$p(\sigma_\varepsilon^2|y, \lambda, \beta, \alpha) = p(\sigma_\varepsilon^2|y, \lambda, \beta) = \mathsf{Inv}\text{-}\chi^2(\sigma_\varepsilon^2|\nu, s^2), \tag{5.3}$$

where $\nu = N$ and $s^2 = S(\lambda, \beta)/N$.

Drawing λ. Regarding the conditional posterior distribution for λ, note

$$p(\lambda|y, \sigma_\varepsilon^2, \beta, \alpha) = p(\lambda|y, \sigma_\varepsilon^2, \beta) \propto \mathsf{N}(y|\lambda X\beta, \sigma_\varepsilon^2 I_N)\, p(\lambda). \tag{5.4}$$

In addition, $\mathsf{N}(y|\lambda X\beta, \sigma_\varepsilon^2 I_N) \propto \mathsf{N}(\lambda|m_\lambda, v_\lambda)$ where

$$m_\lambda = \frac{y^\top (X\beta)}{(X\beta)^\top (X\beta)} \qquad \text{and} \qquad v_\lambda = \frac{\sigma_\varepsilon^2}{(X\beta)^\top (X\beta)}. \tag{5.5}$$

Consequently, the conditional distribution for λ is normally distributed if $p(\lambda)$ is normal. If $p(\lambda)$ is truncated normal, then so is the conditional posterior.

Drawing α. Note

$$p(\alpha|y, \lambda, \sigma_\varepsilon^2, \beta, \alpha) = p(\alpha|\beta) \propto p(\beta|\alpha)\, p(\alpha). \tag{5.6}$$

For making draws from the posterior using a random-walk Metropolis sampler, it is convenient to change variables: let $z = \log(\alpha)$. The likelihood for z is

$$p(\beta|z) = p(\beta|\alpha)|_{\alpha=e^z}, \tag{5.7}$$

where the likelihood for α is given by (4.2). The prior for z is given in (4.3). The Metropolis step proceeds as follows: Given some average step-size s, make random-walk proposals of $z' \sim \mathsf{N}(z, s^2)$. The acceptance ratio is given by

$$\rho(z, z') := \frac{p(\beta|z')\, p(z')}{p(\beta|z)\, p(z)}, \tag{5.8}$$

and the updated value for z is given by

$$z^{(r)} = \begin{cases} z' & \rho(z, z') \geq u \\ z^{(r-1)} & \text{otherwise} \end{cases}, \tag{5.9}$$

where $u \sim \mathsf{Uniform}(0, 1)$.

Drawing β. Drawing β from its conditional posterior distribution is made somewhat complicated by (i) the restriction of β to the simplex, (ii) the non-conjugate prior for β, and (iii) the potential under-identification of β. A "one-at-a-time" Gibbs sampler performs the task well. Owing to the non-conjugate prior, each of the draws is computed via a Metropolis–Hastings step. It is important that the proposal distribution be calibrated to the scale of each coefficient (which varies dramatically across the individual coefficients). I describe the scheme in three stages: First I address a technical detail that involves the simplex; second I describe the proposal distribution; and third I display the Metropolis–Hastings sampler.

Technical detail. Let $\beta_{-j} := \beta \setminus \{\beta_j\}$. The conditional distribution for $\beta_j|\beta_{-j}$ is degenerate because the value of β_j is fixed by β_{-j} (since $\sum_{j=1}^K \beta_j = 1$). By eliminating one of the components, say β_k, we can then cycle through the remaining $K-1$ components via a one-at-a-time Gibbs sampler. However, the magnitude of β_k will affect the efficiency of the sampler. If β_k is close to zero, we will find ourselves back in the previous trap with little or no wiggle room to draw β_j. Therefore, for each sweep of the Gibbs sampler, remove the largest component from the previous sweep and sample over the remaining components.

In order to implement this approach, let

$$\begin{aligned} S^k(\lambda,\beta) &:= S(\lambda,\beta)|_{\beta_k = 1-\sum_{j\neq k}\beta_j} \\ &= (y^k_\lambda - \lambda X^k\beta)^\top (y^k_\lambda - \lambda X^k \beta), \end{aligned} \tag{5.10}$$

where y^k_λ is given in (2.7) and the columns of X^k are given in (2.8). For future reference, note that (for any $j \neq k$)

$$S^k(\lambda,\beta) = c^k_{0j} + c^k_{1j}\,\beta_j + c^k_{2j}\,\beta_j^2, \tag{5.11}$$

where the coefficients $(c^k_{0j}, c^k_{1j}, c^k_{2j})$ are functions of $(X, y, \beta_{-(j,k)}, \lambda)$ and are thus free of β_j and β_k. In particular,[8]

$$c^k_{1j} = -2\,\lambda^2 \left((X^k_{*j})^\top X^k_{*j}\,\beta_j + \lambda^{-1}(X^k_{*j})^\top y^k_\lambda - (X^k_{*j})^\top X^k\beta \right) \tag{5.12a}$$

$$c^k_{2j} = \lambda^2\,(X^k_{*j})^\top X^k_{*j}. \tag{5.12b}$$

Proposal distribution. Given (5.11) and (2.3), the conditional likelihood for β_j (having eliminated β_k) is proportional to a truncated normal distribution:

$$p(y|\lambda,\beta,\sigma^2_\varepsilon) \propto \exp\!\left(\frac{-S^k(\lambda,\beta)}{2\,\sigma^2_\varepsilon}\right) \propto \mathsf{N}_{[0,b^k_j]}(\beta_j|m^k_j, v^k_j), \tag{5.13}$$

where

$$b^k_j = \beta_j + \beta_k = 1 - \sum_{i\neq j,k} \beta_i \tag{5.14}$$

[8] Although both β_j and β_k appear explicitly in (5.12a), each has a zero coefficient when terms are collected. The same statement applies to (5.15a) below.

and [from (5.11) and (5.12)]

$$m_j^k = -\frac{1}{2}\,\frac{c_{1j}^k}{c_{2j}^k} = \beta_j + \frac{\lambda^{-1}(X_{*j}^k)^\top y - (X_{*j}^k)^\top X_{*k} - (X_{*j}^k)^\top X^k \beta}{(X_{*j}^k)^\top X_{*j}^k} \tag{5.15a}$$

$$v_j^k = \frac{\sigma_\varepsilon^2}{c_{2j}^k} = \frac{\sigma_\varepsilon^2}{\lambda^2 (X_{*j}^k)^\top X_{*j}^k}. \tag{5.15b}$$

Two remarks regarding (5.15). First, we require $c_{2j}^k > 0$; this condition is equivalent to $X_{*j} \neq X_{*k}$. Second, in going from (5.12) to (5.15) we have used

$$\lambda^{-1}(X_{*j}^k)^\top y_\lambda^k = \lambda^{-1}(X_{*j}^k)^\top y - (X_{*j}^k)^\top X_{*k}. \tag{5.16}$$

Note that $(X_{*j}^k)^\top X^k$ and $(X_{*j}^k)^\top y$ can be calulated without reference to β or λ.

Drawing β_j. Here I describe the Metropolis–Hastings step. I use the truncated normal distribution in (5.13) to draw the proposal β_j'. In addition, set $\beta_k' = b_j^k - \beta_j'$ and $\beta_\ell' = \beta_\ell$ for $\ell \notin \{j, k\}$. The density for this proposal can be expressed as

$$q_j^k(\beta'|\beta) = \mathsf{N}_{[0,b_j^k]}(\beta_j'|m_j^k, v_j^k). \tag{5.17}$$

Because the proposal density is proportional to the conditional likelihood, the acceptance ratio depends only on the prior ratio:

$$\begin{aligned} \rho_j^k(\beta,\beta') &:= \frac{p(y|\lambda,\beta',\sigma_\varepsilon^2)\,p(\beta'|\alpha,\xi)/q_j^k(\beta'|\beta)}{p(y|\lambda,\beta,\sigma_\varepsilon^2)\,p(\beta|\alpha,\xi)/q_j^k(\beta|\beta')} \\ &= \frac{p(\beta'|\alpha,\xi)}{p(\beta|\alpha,\xi)} = \left(\frac{\beta_j'}{\beta_j}\right)^{\alpha\,\xi_j - 1} \left(\frac{\beta_k'}{\beta_k}\right)^{\alpha\,\xi_k - 1}. \end{aligned} \tag{5.18}$$

Let β° denote the state of β just prior to the update for β_j during the current "sweep" of the sampler (sweep r). Note that β° may contain some elements that have already been updated during the current sweep and other elements that have not. In particular, $\beta_j^\circ = \beta_j^{(r-1)}$. Then the updated value for β_j is given by

$$\beta_j^{(r)} = \begin{cases} \beta_j' & \rho_j^k(\beta^\circ, \beta') \geq u \\ \beta_j^{(r-1)} & \text{otherwise} \end{cases}, \tag{5.19}$$

where $u \sim \mathsf{Uniform}(0,1)$.

Note that if the prior for β were flat, then $\alpha\,\xi_j = \alpha\,\xi_k = 1$. In this case $\rho_j^k(\beta,\beta') \equiv 1$ and the proposal would always be accepted. This is a case where the Metropolis–Hastings sampler reduces to the Gibbs sampler.

Alternative approach to drawing β. This section describes a Metropolis–Hastings sampler for β that involves reparametrizing β. This sampler is much easier to implement and may be nearly as efficient as the previously described sampler for β.

Let $v = (v_1, \ldots, v_K) \in \mathbb{R}^K$. Further, let the prior distribution for v be given by $p(v) = \prod_{i=1}^K p(v_j)$, where

$$p(v_j) = \mathsf{ExpGamma}(v_j|a_j, 1) = \frac{e^{a_j\,v_j - e^{v_j}}}{\Gamma(a_j)}. \tag{5.20}$$

By definition, $e^{v_j} \sim \mathsf{Gamma}(a_j, 1)$.

Express β in terms of v:

$$\beta = \frac{e^v}{\sum_{j=1}^K e^{v_j}}. \tag{5.21}$$

Note $\beta \sim \mathsf{Dirichlet}(a)$, where $a = (a_1, \ldots, a_K)$. The parameter for the Dirichlet distribution can be re-expressed as $a = \alpha\,\xi$ where $\alpha = \sum_{j=1}^K a_j$ and $\xi = a/\alpha$. Also note the scaling relation between v and β:

$$\frac{d\beta_k}{dv_j} = \begin{cases} \beta_j\,(1-\beta_j) & j = k \\ -\beta_j\,\beta_k & j \neq k \end{cases}. \tag{5.22}$$

The Metropolis–Hastings sampling strategy for v_j involves the following proposal:

$$q(v_j'|v_j) = \mathsf{N}\big(v_j'|v_j, \varsigma(v_j)^2\big), \tag{5.23}$$

where $\varsigma(v_j)$ is a step-size function. The step-size function must be determined empirically. As an example,

$$\varsigma(v_j) = a\left(\frac{1}{2} - \tan^{-1}\left(\frac{v_j - m}{s}\right)/\pi\right), \tag{5.24}$$

for suitable (a, m, s).[9] Because this scale factor depends on v_j, a "Hastings" correction is required to account for asymmetry between $q(v_j'|v_j^{(r-1)})$ and $q(v_j^{(r-1)}|v_j')$.

Let[10]

$$v^0 = (v_1, \ldots, v_j^{(r-1)}, \ldots, v_K) \tag{5.25}$$

$$v^1 = (v_1, \ldots, v_j', \ldots, v_K) \tag{5.26}$$

and let β^0 and β^1 be computed from v^0 and v^1. Then

$$v_j^{(r)} = \begin{cases} v_j' & \mathcal{M} \geq u \\ v_j^{(r-1)} & \text{otherwise} \end{cases}, \tag{5.27}$$

where $u \sim \mathsf{Uniform}(0, 1)$ and

$$\mathcal{M} = \frac{p(y|\lambda, \beta^1, \sigma_\varepsilon^2)\,p(v_j')/q(v_j'|v_j^{(r-1)})}{p(y|\lambda, \beta^0, \sigma_\varepsilon^2)\,p(v_j^{(r-1)})/q(v_j^{(r-1)}|v_j')}. \tag{5.28}$$

6. Second posterior sampler

This section describes an importance sampler. This sampler is extremely simple to implement but it can be so inefficient as to be useless. Nevertheless, there are cases where it can be useful. For example, a variant of this sampler is used in Fisher (2015).

[9] Preliminary testing suggests that $a = -(K/\alpha)$, $m = -\frac{1}{2}(K/\alpha)$, and $s = \frac{1}{2}(K/\alpha)\tan\Big(\frac{\pi/10}{K/\alpha}\Big)$ works reasonably well.

[10] Some of the other components of v may have already been updated to (r) while others may not have been.

Begin by integrating out σ_ε^2 analytically, producing the likelihood for (λ, β):

$$p(y|\lambda, \beta) = \int_0^\infty \frac{\mathsf{N}(y|\lambda X\beta, \sigma_\varepsilon^2 I_N)}{\sigma_\varepsilon^2} \, d\sigma_\varepsilon^2 \propto S(\lambda, \beta)^{-N/2}, \tag{6.1}$$

where $S(\lambda, \beta)$ is given in (2.4). Assume λ has a proper prior or is fixed.

Draw $\{\beta^{(r)}\}_{r=1}^R$ and $\{\lambda^{(r)}\}_{r=1}^R$ from their priors and evaluate the likelihood for each draw [see (6.1)]:

$$q^{(r)} := S(\lambda^{(r)}, \beta^{(r)})^{-N/2}. \tag{6.2}$$

It is convenient to define

$$\widehat{L} := \sum_{r=1}^R q^{(r)} \qquad \text{and} \qquad \widetilde{\beta} := \sum_{r=1}^R q^{(r)} \, \beta^{(r)}. \tag{6.3}$$

Estimates of the marginal likelihood of the model and the posterior expectation of β are given by

$$\widehat{\ell} := \widehat{L}/R \qquad \text{and} \qquad \overline{\beta} := \widetilde{\beta}/\widehat{L}. \tag{6.4}$$

It is easy to parallelize this sampler using a number of *batches* with R^b draws in batch b. For each batch compute $\widehat{L}^b$ and $\widetilde{\beta}^b$ according to (6.3). Then

$$R = \sum_{b=1}^{\mathcal{B}} R^b, \qquad \widehat{L} = \sum_{b=1}^{\mathcal{B}} \widehat{L}^b, \qquad \text{and} \qquad \widetilde{\beta} = \sum_{b=1}^{\mathcal{B}} \widetilde{\beta}^b, \tag{6.5}$$

where $\mathcal{B}$ is the number of batches.

Appendix A. A class of basis densities and distributions

Basis densities. One class of basis densities is the Bernstein-Skew class,[11] a special case of which is composed of Beta-Normal distributions. Let $\Phi(\,\cdot\,)$ denote the cumulative distribution function (CDF) for the standard normal distribution. The density for the Beta distribution is given by

$$\mathsf{Beta}(z|a, b) = z^{a-1} \, (1-z)^{b-1}/B(a, b), \tag{A.1}$$

where the beta function is $B(a, b) = \int_0^1 z^{a-1} \, (1-z)^{b-1} \, dz$. Then the Beta-Normal basis densities are given by

$$f_j(x) = \mathsf{Beta}\Big(\Phi\Big(\frac{x-\mu}{\eta}\Big) \,\Big|\, j, K-j+1\Big) \, \mathsf{N}(x|\mu, \eta^2). \tag{A.2}$$

These basis densities are related to Bernstein polynomials from which they inherit the following adding-up property:

$$\frac{1}{K} \sum_{j=1}^K f_j(x) = \mathsf{N}(x|\mu, \eta^2). \tag{A.3}$$

With this in mind, these basis densities may be interpreted as providing nonparametric variation around a base distribution (which in this case is the normal distribution).

[11] See Quintana et al. (2009).

Basis distributions. For the second example, it is convenient to use basis distributions. For basis distributions, define

$$F_j(x) := I_{\Phi\left(\frac{x-\mu}{\eta}\right)}(j, K-j+1), \tag{A.4}$$

where $I_z(a,b)$ is the regularized incomplete beta function (i.e., the CDF of the Beta distribution). Note that $F_j'(x) = f_j(x)$ as given in (A.2).

References

Aït-Sahalia, Y. and J. Duarte (2003). Nonparametric option pricing under shape restrictions. *Journal of Econometrics 116*, 9–47.

Carlson, J. B., B. R. Craig, and W. R. Mellick (2005). Recovering market expectations of FOMC rate changes with options on federal funds futures. *Journal of Futures Markets 25*, 1203–1242.

Fisher, M. (2015). Fitting a distribution to survey data for the half-life of deviations from PPP. Working Paper 2015-15, Federal Reserve Bank of Atlanta.

Quintana, F. A., M. F. J. Steel, and J. T. A. S. Ferreira (2009). Flexible univariate continous distributions. *Bayesian Analysis 4*(4), 497–522.

Federal Reserve Bank of Atlanta, Research Department, 1000 Peachtree Street N.E., Atlanta, GA 30309–4470

E-mail address: mark.fisher@atl.frb.org

URL: http://www.markfisher.net