Predictive Modeling with the Tweedie Distribution

Glenn Meyers
ISO Innovative Analytics

CAS Annual Meeting – Session C-25
November 16,2009

Background – The Collective Risk Model

Describe as a simulation algorithm

1. Select a random number of claims, *N*
2. For *i* = 1 to *N*
   - Select a random claim amount, Z_i
3. Total Loss = $\sum_{i=1}^{N} Z_i$

History
1980 Discussion Paper Program

- Glenn Meyers
  - Used collective risk model simulation for retrospective rating.
- Shaw Mong
  - Used Fourier transforms to calculate collective risk model probabilities ***WITHOUT*** simulation.
  - Assumed that claim severity had a gamma distribution.
  - Later generalized by Heckman and Meyers (1983)

History – Statistical Community (1984)

- University of Iowa Department of Statistics
  - Poisson frequency, gamma severity
  - Used in order restricted inference
- Tweedie publishes paper
  - Tweedie, M. C. K., “An Index which Distinguishes between Some Important Exponential Families,” in *Statistics: Applications and New Directions, Proceedings of the Indian* Statistical Golden Jubilee International Conference, J. K. Ghosh and J. Roy (Eds.), Indian Statistical Institute, 1984, 579—604.

Uses of Collective Risk Model

- Calculating Aggregate Loss Distributions
  - Retrospective rating
  - Reinsurance
  - Enterprise risk management
- Fitting models of insurance loss data
  - Simulation is of no help in maximum likelihood estimation
  - The Tweedie distribution is a member of the exponential dispersion family and thus can be used in a GLM

Tweedie as a Compound Poisson Model

- Claim Count *N* ~ Poisson(λ)
- Claim Severity *Z* ~ Gamma(α,θ)
  - KPW *Loss Models* parameters
- Translate into standard Tweedie parameters

$$p=\frac{\alpha+2}{\alpha+1},\;\; \mu=\lambda\cdot\alpha\cdot\theta,\;\; \phi=\frac{\lambda^{1-p}\cdot\left(\alpha\cdot\theta\right)^{2-p}}{2-p}$$

□ With some algebra we can see that

$$Var\left[Y\right]=\phi\cdot\mu^{p}=\lambda\cdot\theta^{2}\cdot\alpha\cdot\left(\alpha+1\right)$$

- This is the same as predicted by well known collective risk model variance formulas

Interpreting the Tweedie “p”

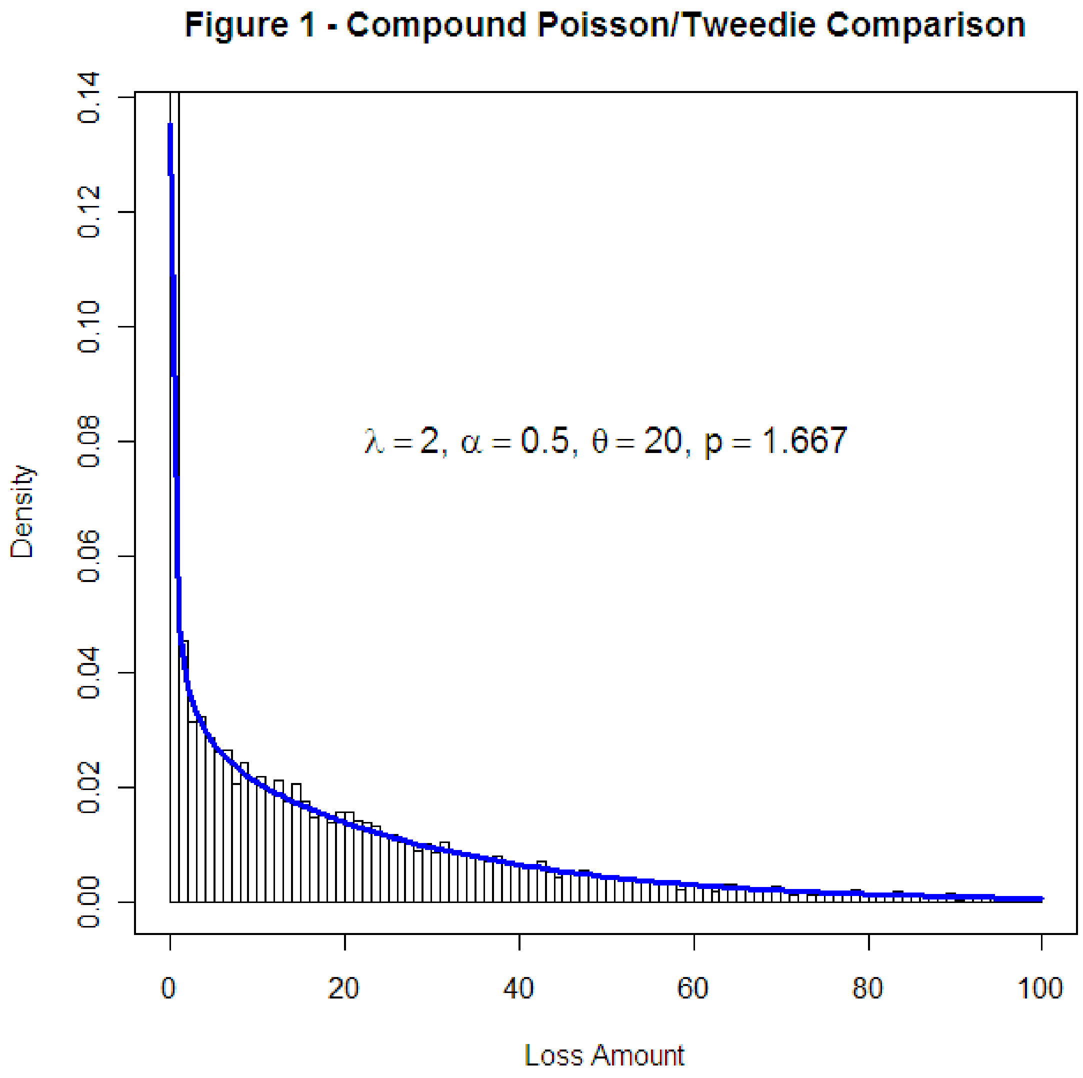
$p = \frac{\alpha + 2}{\alpha + 1}$, CV for gamma distribution = $\frac{1}{\sqrt{\alpha}}$

As $\alpha \to \infty$, CV $\to 0$ and $p \to 1$

- For $p = 1$, the Tweedie is called the Overdispersed Poisson distribution
  - Claim amounts are constant
- For most P/C insurance applications
  - CV > 1 which implies $p > 1.5$

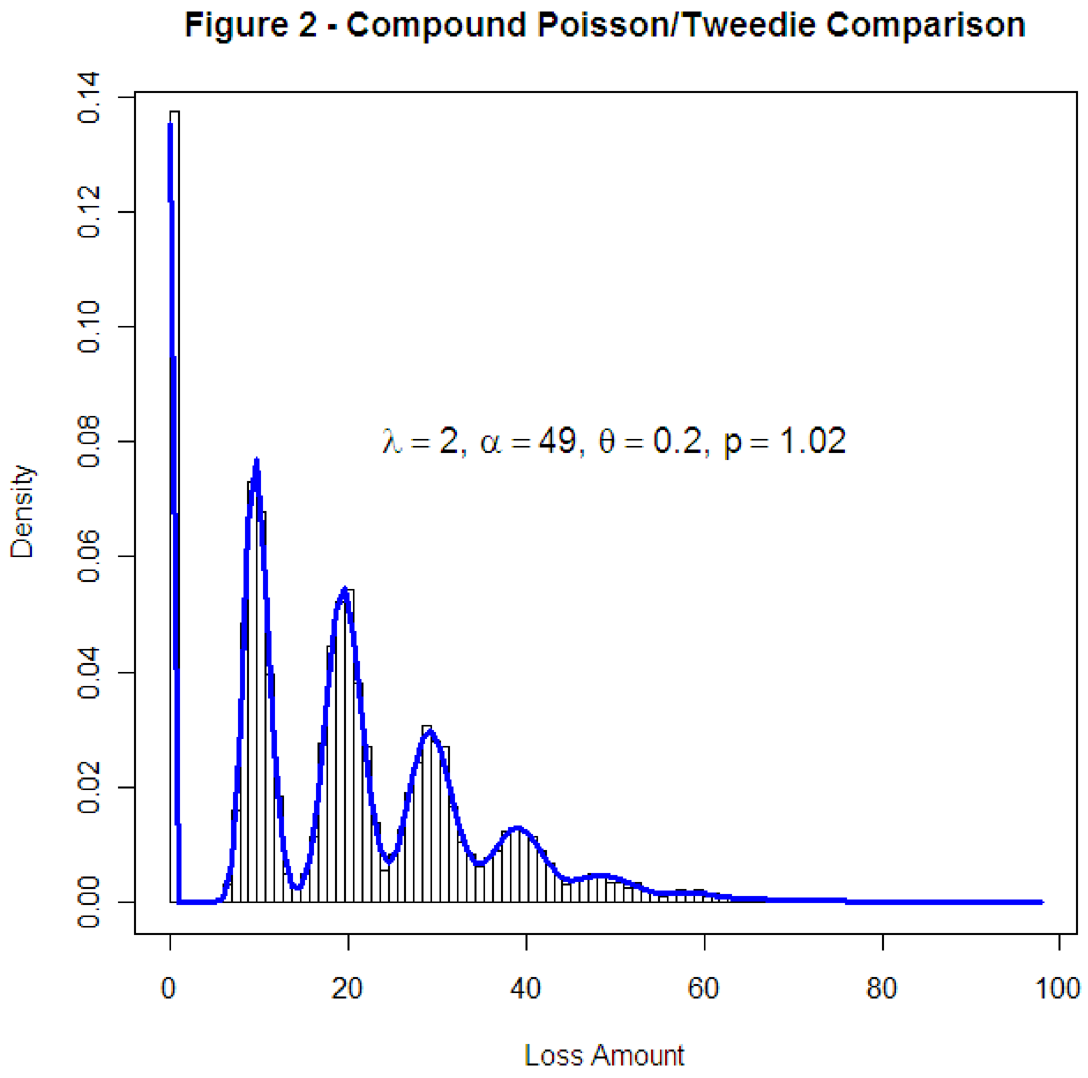
Illustrating the Effect of *p*

Simulation vs R's "tweedie" package



Illustrating the Effect of *p*

Simulation vs R's "tweedie" package



Tweedie Density Function

- Dispersion model form – Dunn and Smyth(2008)

$$f\left(y \mid p,\mu,\phi\right)=f\left(y \mid p,y,\phi\right)\cdot\exp\left(-\frac{1}{2\phi}d\left(y,\mu\right)\right)$$

- Deviance - $d(y,\mu)$

$$d\left(y,\mu\right)=2\cdot\left(\frac{y^{2-p}}{\left(1-p\right)\cdot\left(2-p\right)}-\frac{y\cdot\mu^{1-p}}{1-p}+\frac{\mu^{2-p}}{2-p}\right)$$

GLMs with the Tweedie Distribution

- Maximize log-likelihood   Minimize Deviance
- GLMs focus only on estimating μ
  - p and ϕ are either given, or estimated outside the GLM framework.
- Unnecessary to evaluate $f(y|p,y,\phi)$
  - Very fortunate for GLM
- Not helpful for more general models
  - Dunn and Smyth (2005,2008) evaluate $f(y|p,y,\phi)$ using complicated math involving series expansions and Fourier inversion. It is also computationally slow.
  - Dunn is the author of the Tweedie package in R.

Problem with the Compound Poisson Interpretation of the Tweedie Distribution

$$\phi = \frac{\lambda^{1-p} \cdot (\alpha\theta)^{2-p}}{2-p} = \frac{\mu^{2-p}}{\lambda \cdot (2-p)} = \frac{\alpha\theta \cdot \mu^{1-p}}{2-p}$$

- A constant ϕ will force an artificial relationship between the claim frequency, λ, or the claim severity, $\alpha\theta$.
- Uses of Tweedie distribution
  - Desire to build pure premium models where claim frequency and claim severity have their own independent variables.
  - Monte-Carlo Markov Chain simulations
    - Need speed

Rearrange Calculation of Tweedie Density

- Objectives of method
  - Allow ϕ to vary as an input
  - Computationally fast
- Keep *p* fixed
  - Current applications can reliably estimate *p.*
  - Experience indicates that pure premium estimates are relatively robust with respect to *p.*
  - Be careful if accurate estimates of frequency are required.

Tweedie Density Function

- Dispersion model form – Dunn and Smyth(2008)

$$f\left(y \mid p,\mu,\phi\right)=f\left(y \mid p,y,\phi\right)\cdot \exp\left(-\frac{1}{2\phi}d\left(y,\mu\right)\right)$$

- First term - $f(y|p,y,\phi)$
  - Slow to calculate when called
  - Not any slower for calling with a long vector *y*

Dunn-Smyth Rescaling (2008)

$$f\left(y\,|\,p,\mu,\phi\right)=k\cdot f\left(ky\,|\,p,k\mu,k^{2-p}\phi\right)$$

- For given λ, θ, α use the above formulas to calculate ϕ, μ and *p*.
- Find a short and fast approximation for $f(y|p,y,1)$
- Given *f*, calculate *k* and use Dunn-Smyth rescaling and the approximation to calculate for $y > 0$:

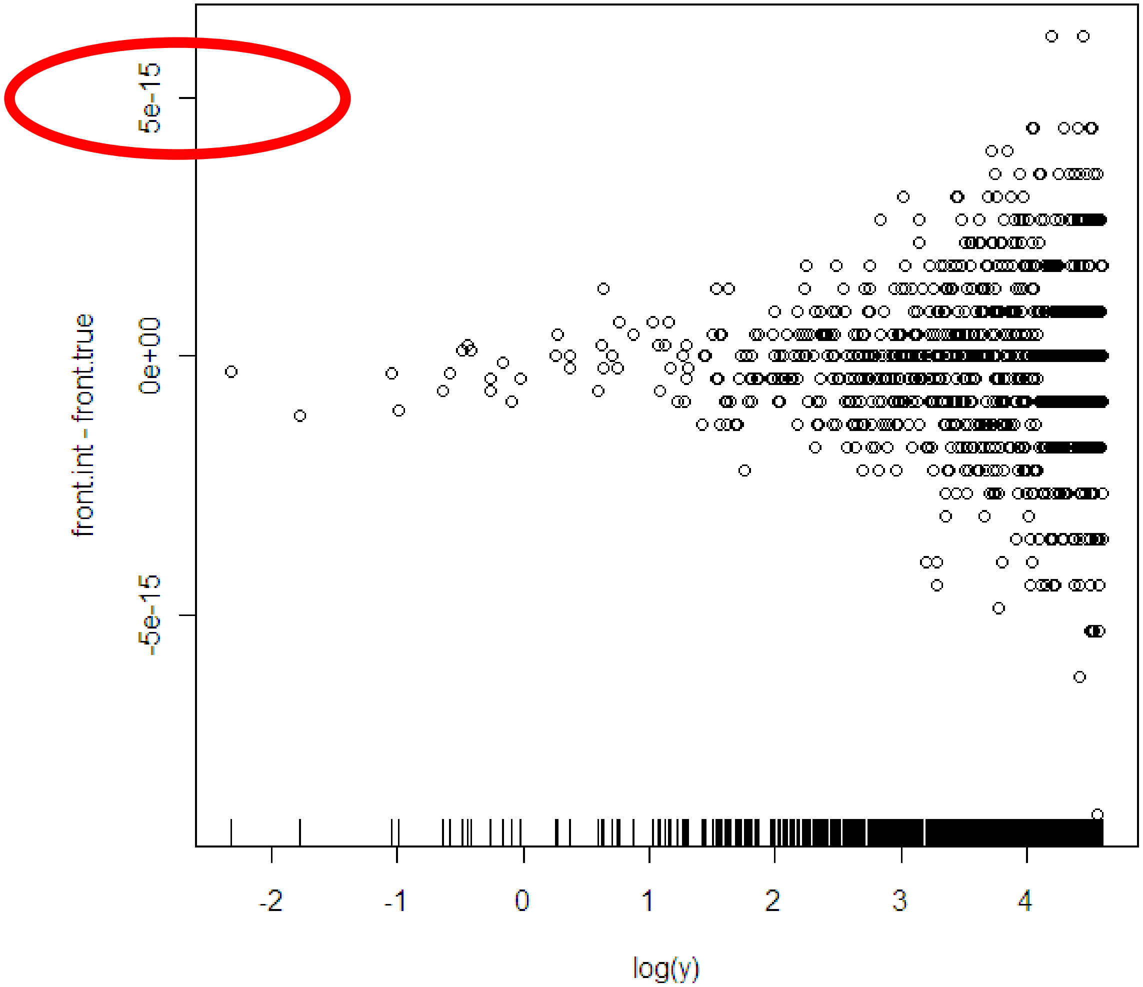
Set $k=\phi^{-\frac{1}{2-p}}$, then $f\left(y\,|\,p,y,\phi\right)=k\cdot f\left(ky\,|\,p,ky,1\right)$

- For $y = 0$, set $f\left(0\,|\,p,\mu,\phi\right)=\exp\left(-\dfrac{\mu^{2-p}}{\phi\cdot\left(2-p\right)}\right)$

Approximating $\log(f(y|p,y,1))$

- Calculate, $\log(f(y|p,y,1))$ using “dtweedie”, ***ONCE*** with a long vector y_{fixed}.
- Find 3 values in the vector y_{fixed} that are close to y.
- Use divided differences interpolation to approximate $\log(f(y|p,y,1))$.
- Increase accuracy by putting more points in y_{fixed}.

Accuracy with 10,000 y_{fixed}s



R Code for Last Plot

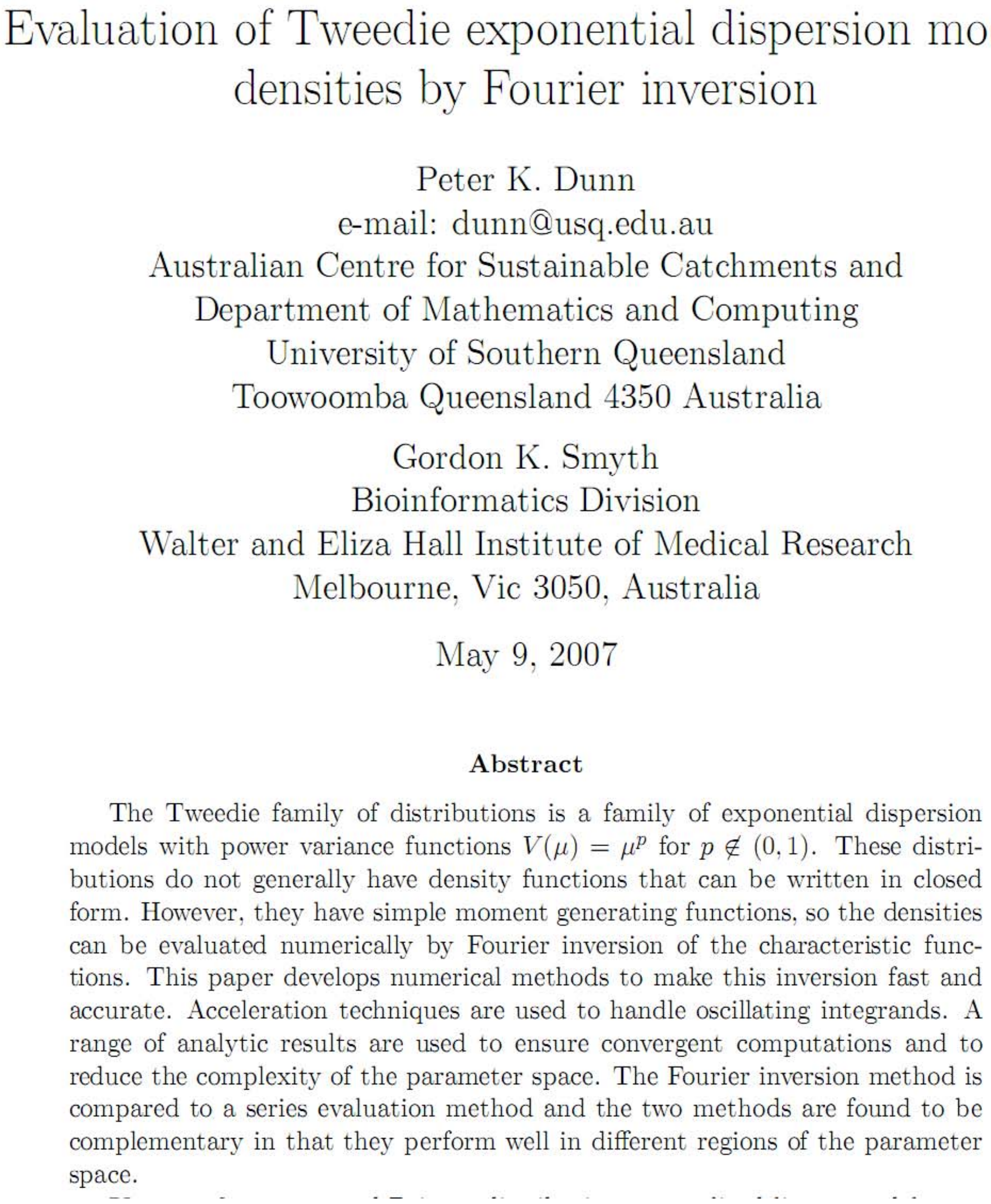
```
library(statmod)
library(tweedie)
p=1.5
num=10000
log.ybot=log(.01)
log.ytop=log(100)
del=(log.ytop-log.ybot)/num
log.y1=seq(from=log.ybot,to=log.ytop,length=num)
front=dtweedie(exp(log.y1),p,exp(log.y1),1)
#
ldtweedie.front=function(y,lyf,front){
 lf=log(front)
 ly=log(y)
 del=lyf[2]-lyf[1]
 low=pmax(floor((ly-lyf[1])/del),1)
 d01=(lf[low+1]-lf[low])/del
 d12=(lf[low+2]-lf[low+1])/del
 d23=(lf[low+3]-lf[low+2])/del
 d012=(d12-d01)/2/del
 d123=(d23-d12)/2/del
 d0123=(d123-d012)/3/del
 ld=lf[low]+(ly-lyf[low])*d01+(ly-lyf[low])*(ly-lyf[low+1])*d012+
        (ly-lyf[low])*(ly-lyf[low+1])*(ly-lyf[low+2])*d0123
 return(ld)
 }
#
```

```
ldtweedie.scaled=function(y,p,mu,phi){
 dev=y
 ll=y
 k=(1/phi)^(1/(2-p))
 ky=k*y
 yp=ky>0
 dev[yp]=2*((k[yp]*y[yp])^(2-p)/((1-p)*(2-p))-k[yp]*y[yp]*
   (k[yp]*mu[yp])^(1-p)/(1-p)+(k[yp]*mu[yp])^(2-p)/(2-p))
 ll[yp]=log(k[yp])+ldtweedie.front(ky[yp],log.y1,front)-dev[yp]/2
 ll[!yp]=-mu[!yp]^(2-p)/phi[!yp]/(2-p)
 return(ll)
 }
#
runif(1000,min=0,max=exp(log.ytop))
front.true=log(dtweedie(y,p,y,1))
front.int=ldtweedie.front(y,log.y1,front)
plot(log(y),front.int-front.true)
rug(log(y))
summary(front.int-front.true)
```

Remarks

- $f(y|p,y,\phi)$ and $\{\alpha_i\}$ can be calculated in R with the tweedie package.
- log of Tweedie densities can quickly calculated in closed form using the cubic approximation.
  - Fast calculation of log-likelihood is necessary for Bayesian estimation using the MCMC simulations.
  - Coefficients of the cubic approximation allow for easy coding in SAS and Excel for MLE estimation of ϕ.

Dunn, P. K. & Smyth, G. K. (2008). Evaluation of Tweedie exponential dispersion model densities by Fourier inversion. *Statistics and Computing*, **18**, 73–86.



Evaluation of Tweedie exponential dispersion model densities by Fourier inversion

Peter K. Dunn
e-mail: dunn@usq.edu.au
Australian Centre for Sustainable Catchments and
Department of Mathematics and Computing
University of Southern Queensland
Toowoomba Queensland 4350 Australia

Gordon K. Smyth
Bioinformatics Division
Walter and Eliza Hall Institute of Medical Research
Melbourne, Vic 3050, Australia

May 9, 2007

Abstract

The Tweedie family of distributions is a family of exponential dispersion models with power variance functions $V(\mu) = \mu^p$ for $p \notin (0, 1)$. These distributions do not generally have density functions that can be written in closed form. However, they have simple moment generating functions, so the densities can be evaluated numerically by Fourier inversion of the characteristic functions. This paper develops numerical methods to make this inversion fast and accurate. Acceleration techniques are used to handle oscillating integrands. A range of analytic results are used to ensure convergent computations and to reduce the complexity of the parameter space. The Fourier inversion method is compared to a series evaluation method and the two methods are found to be complementary in that they perform well in different regions of the parameter space.